THE UNIVERSITY *of* EDINBURGH

Machine learning for retinal image analysis

Justin Engelmann

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2024

Für meine Familie, die unentwegt an mich glaubt.
Für Api, den ich immer in meinem Herzen tragen werde.

In die Erd ist's aufgenommen,
Glücklich ist die Form gefüllt,
Wird's auch schön zutage kommen,
Daß es Fleiß und Kunst vergilt?
Wenn der Guß mißlang?
Wenn die Form zersprang?
Ach! vielleicht indem wir hoffen,
Hat uns Unheil schon getroffen.

Abstract

Retinal images, images of the retina at the back of our eyes, are an important part of modern ophthalmology and further capture the retinal vasculature and nerves, which could allow insight into cardio- and neurovascular disease. This is especially promising as retinal images are non-invasive, fast-to-acquire and low-cost compared to other types of medical imaging such as brain magnetic resonance imaging. A variety of retinal imaging modalities exist, most importantly traditional colour fundus photography (CFP) and optical coherence tomography (OCT). CFP is the most widespread type of retinal imaging and captures a true colour en-face image of the retina, typically with a field of view of around 45 degrees. OCT imaging captures the retina in depth and thus allows assessment of individual layers of the retina and – with modern methods such as Enhanced Depth Imaging – even captures the choroid, a dense vascular tissue beneath the retina. More recent modalities include OCT angiography which uses repeated OCT images to estimate blood flow and ultra-widefield fundus imaging which captures most of the retina with a field of view of around 200 degrees. Retinal imaging is already widespread and continuously proliferating: lower-cost handheld devices or smartphone addons make CFP available in lower resource settings, while once cutting-edge OCT can now be found at high-street opticians in the UK.

Retinal images provide a wealth of information but are complex to analyse, in part due to variations in image quality, anatomy, and retinal pathology that make traditional development of handcrafted analysis pipelines challenging. The recent decade saw great advances in machine learning methods, particularly deep learning for computer vision. Instead of manually designing a pipeline, a machine learning model is a parameterised pipeline that can be fit to training data to approximate the mapping from inputs to outputs. This approach is highly effective for many vision tasks, including classification, regression and segmentation.

In this thesis, I present three themes of work using machine learning for retinal image analysis. First, using machine learning for retinal disease detection. Second, using machine learning for developing efficient and robust automated analysis pipelines for retinal imaging. And third, validating and applying these tools.

For the first theme, I developed a deep learning model that can detect seven key retinal diseases in ultra-widefield pseudo-colour retinal images with very promising performance and investigate which regions of the ultra-widefield images are important for automated disease detection in a data driven way.

For the second theme, I developed three tools. First, deep approximation of retinal traits, or DART for short, that computes retinal fractal dimension (FD), a metric relating to the complexity of the blood vessels in CFPs, orders of magnitudes faster and more robustly than traditional methods. Second, jointly with a colleague, I developed a tool initially for segmenting the choroid region in OCT, called DeepGPET. Next, we developed Choroidalyzer, which segments the choroid and the choroidal vasculature while also identifying the location of the fovea. This allows for fully-automated computation of choroidal thickness, area, vascular index in a fovea-centred region of interest. Third, I developed QuickQual an efficient and easy-to-use method for CFP quality assessment that obtains state-of-the-art performance on a commonly used quality assessment dataset.

Finally, for the third theme, I applied DART to real-world, primary care data and found a significant association between lower FD and prevalent systemic health conditions. Furthermore, I compared the repeatability and robustness of DART to AutoMorph, a method that follows the traditional paradigm for computing FD, finding that DART was not only more robust to image quality issues but also more repeatable even for high quality images.

In my opinion, this thesis exemplifies the potential of machine learning for retinal image analysis. I hope that my work will – eventually and incrementally – advance the field of retinal image analysis and one day make a positive difference for clinical practice.

Lay Summary

I used computers to process pictures of the eye. Those images show the “retina”. The retina sits at the back of our eyes. It allows us to see. The retina processes light, like a camera sensor. Thus, the retina is very important. If someone’s retina is unhealthy, they might have poor vision or even become blind.

Computers that find unhealthy retinas would be very useful. Optometrists can take these pictures. Optometrists check eyes, but are not doctors. Even some opticians can take them. Opticians make glasses. Imagine if they had a computer to look for problems. If they find a problem, the person can then see an eye doctor. That doctor can then treat them so they do not become blind.

To create such computers, I used “machine learning”. This is an approach that allows the computer to learn from data. For example, images of retinas and whether they are healthy or not. With that, the computer can learn to find unhealthy eyes. This is what I did in my thesis.

You might have heard of “AI”. That is the same thing as machine learning. I used machine learning because it works very well. I can develop a very accurate computer.

In the eye pictures, we can see the “blood vessels”. Blood vessels are like very small pipes through which the blood flows. Healthy blood vessels are important for your eyes. But they are also important your whole body. For example, your heart or your brain.

We can easily take images of your eye to see the blood vessels. Taking images of your heart or brain is very difficult. They need very expensive machines. Eye images are very quick and painless. Brain and heart images are slow and not very pleasant.

So, I taught computers to look at blood vessels in eye pictures. I taught the computer to be very accurate. Now, we can find very small changes. I hope that this could one day be used to check people. Are they about to have a heart attack or stroke? Then, they can get help from a doctor before the problem occurs.

I really hope that my work will one day help people like you and me and our families. It could keep our eyes healthy. And maybe our hearts and brains, too. Other scientists might also use my work to do exciting research.

Acknowledgements

I thank my family and partner for their love and unwavering support, for always helping me grow and yet accepting me for who I currently am.

I thank my supervisors for their mentorship and support, not only with my PhD but with starting my career. I am certainly quite demanding, yet even I could not have asked for better supervisors.

I thank all my educators who laid the foundations that allowed me to walk this path. I especially thank my teachers, mentors, and headmaster at St. Afra, to whom I owe my education. I will be forever grateful to you for your heroic efforts. I aspire to make the most of the gift you have given me.

I thank my colleagues, collaborators, and friends across the globe who inspire me with their curiosity, passion, and brilliance.

Declaration

I declare that this thesis was composed by myself, that the work contained herein - including the publications - is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Justin Engelmann

Contents

Chapter 1

Introduction

1.1 Motivation & core themes

The overarching motivation for my work is to apply machine learning to retinal image analysis where it is useful and might one day make a positive impact to people's lives. From that, I derive three core themes for this work. First, using machine learning for retinal disease detection in retinal imaging. Second, using machine learning to develop retinal image analysis tools that are useful to researchers and might eventually even find clinical use cases. Third, validating and applying these tools. In my opinion, each of these themes has the potential to contribute towards my overarching motivation.

Retinal disease detection is already an important part of ophthalmic care and automated methods could be useful here. For example, for disease screening in settings where no human expert is available, or to provide clinical decision support by highlighting areas of potential pathology and providing a virtual second opinion. In addition to being useful, retinal disease detection from retinal images is also a task that is known to be feasible: while additional information is considered in ophthalmic examinations (e.g. symptoms, family history, risk factors, other types of tests and examinations such as visual fields), in many cases the presence of retinal disease can be judged from a retinal image alone. This is in contrast with using retinal imaging for risk prediction for systemic health conditions, where it is not yet obvious that there are meaningful increases in predictive power over simple known risk factors to be had. From a technical perspective, methods for training image classifiers are very mature and easy to apply nowadays. However, developing machine learning models for ophthalmology is a highly interdisciplinary endeavour and while modern yet simple machine learning models are highly effective, appropriately framing the problem and then arriving at an effective solution that could be clinically useful is not necessarily trivial. This theme will be exemplified by Chapter 2 of this thesis which presents a model for detecting several different retinal diseases in ultra-widefield fundus images.

Retinal imaging contains a wealth of information and there are many potential applications for quantitative retinal traits, which capture specific aspects of the images, e.g. relating to vasculature, nerves, layers of the retina, or pathology. These retinal traits could be used for biomedical research, for instance to examine relationship between the eye and other organs, such as vascular changes in the eye, brain, and heart. They are also useful to ophthalmic research specifically, for example by providing an objective measure of disease severity to study the effectiveness of clinical interventions. Finally, these retinal traits might one day find application in risk prediction models for systemic disease. Retinal image analysis is a field with a rich, decades-long tradition and has produced many effective tools that were primarily developed with traditional computer vision methods. Modern machine learning techniques could augment existing approaches to yield more efficient and robust tools. This theme will be exemplified by Chapters 3, 6, and 7 of this thesis. Chapter 3 presents a model for computing retinal fractal dimension, a trait relating to vessel complexity, from colour fundus images in a way that is more robust and orders of magnitudes more efficient than traditional approaches. Chapter 6 presents two models for analysing the choroid in optical coherence tomography imaging. Chapter 7 presents an easy-to-use and efficient model for assessing the quality of colour fundus images.

Once developed, these tools can only be useful if they are indeed used. This necessitates both further validation beyond the results presented in the work introducing the tool as well as applications that demonstrate its value and increase its visibility in the field. Furthermore, developing good tools requires being familiar with the challenges and pain points of the eventual user. This is best accomplished by using the tools oneself, which helps identifying both strengths and shortcomings, a practice the tech industry colourfully calls “dogfooding”. Finally, work that is more focused on discovery than engineering might satisfy the researcher’s curiosity and increase the motivation for developing and refining new tools. This theme is exemplified by Chapter 4 which applies the tool from Chapter 3 to investigate potential associations between retinal fractal dimension and systemic health in real-world clinical data, as well as Chapter 5 which compares the repeatability and robustness of this tool with an approach following a more traditional paradigm. As is the case with all three themes, but especially this one, there is a good number of relevant ongoing work that did not make its way into this thesis but hopefully will constitute additional examples in the near future.

Secondary motivations for the work presented in this thesis are to develop and demonstrate good command of modern machine learning methods applied to medical imaging, to develop a reasonable understanding of retinal imaging, retinal disease, and ophthalmic care, and to build collaborations and relationships for a potential future academic career.

1.2 What's in this thesis

Each proper chapter of this thesis contains at least one peer-reviewed paper, preceded by a short introduction and followed by a short conclusion. In those conclusions, I briefly reflect on the work, highlight some weaknesses and limitations, and then give an outlook regarding future work, some of which is already in progress. My goal is not to restate every single limitation already mentioned in the respective paper itself, so I kindly ask the reader to consider the limitation sections of each paper itself, too. Likewise, in the introduction, I do not intend to rewrite the first section of each paper, but to provide some context and explain my motivation for the given piece of work.

While I prefer the authoritative and objective sounding "we" when writing papers, I want to offer some of my own reflections and views in the sections written specifically for this thesis, which I will try to signpost with the first person. Good science is careful, nuanced, and expresses views supported by evidence, yet I think most researchers have additional thoughts about their work. Such thoughts might not fit in a paper but would be offered when discussing their work with colleagues, for example during poster presentations. I hope that expressing them in this thesis might be of interest to the reader and in my opinion they provide more context to the work presented here than merely repeating the introduction and conclusion sections of the paper itself would.

Chapter 2 presents a machine learning model for retinal disease detection in ultra-widefield fundus images and further looks at what regions of these images were important for model performance. This work (Engelmann et al., 2022a) has been published in Nature Machine Intelligence as:

> Engelmann, J., McTrusty, A. D., MacCormick, I. J., Pead, E., Storkey, A., & Bernabeu, M. O. (2022). Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. Nature Machine Intelligence, 4(12), 1143-1154.

Please note that this work originated as my thesis for the MSc by Research stage of my PhD programme. For the paper, two major changes were made that I worked on during the PhD stage of my programme. First, the work changed substantially in style and length, transforming it from a comprehensive MSc thesis into a more concise research paper. Second, the analysis underwent substantial refinement through the feedback from clinical collaborators whom we involved as co-authors prior to submission of the manuscript and through the feedback of two anonymous peer reviewers. Furthermore, said MSc by Research was a mandatory part of the very same PhD programme the present thesis is submitted for. I am thus of the opinion that this work is relevant to the present thesis.

Chapter 3 presents a machine learning model for computing retinal fractal dimension, a measure of vessel complexity, from colour fundus images in an efficient and robust manner. I call this method deep approximation of retinal traits, or DART for short. This work (Engelmann et al., 2022b) has been published in the proceedings of the workshop on Ophthalmic Medical Image Analysis at the 2022 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) as:

> Engelmann, J., Villaplana-Velasco, A., Storkey, A., & Bernabeu, M. O. (2022). Robust and Efficient Computation of Retinal Fractal Dimension Through Deep Approximation. In: Antony, B., Fu, H., Lee, C.S., MacGillivray, T., Xu, Y., Zheng, Y. (eds) Ophthalmic Medical Image Analysis. OMIA 2022. Lecture Notes in Computer Science, vol 13576.

Chapter 4 applies DART to a real-world, primary care dataset from Glasgow to see if retinal fractal dimension is associated with systemic health in data that was not specifically collected for research. It (Engelmann et al., 2024b) has been published in the Association for Research in Vision and Ophthalmology (ARVO) journal Translational Vision Science & Technology (TVST):

> Engelmann, J., Kearney, S., McTrusty, A., McKinlay, G., Bernabeu, M. O., & Strang, N. (2024). Retinal Fractal Dimension Is a Potential Biomarker for Systemic Health—Evidence From a Mixed-Age, Primary-Care Population. Translational Vision Science & Technology, 13(4), 19.

Chapter 5 examines the repeatability and robustness to image quality issues of DART and a pipeline for computing fractal dimension that follows a more traditional paradigm. It (Engelmann et al., 2024c) has been published in the ARVO journal Investigative Ophthalmology & Visual Science (IOVS):

> Engelmann, J., Moukaddem, D., Gago, L., Strang, N., & Bernabeu, M. O. (2024). Applicability of Oculomics for Individual Risk Prediction: Repeatability and Robustness of Retinal Fractal Dimension Using DART and AutoMorph. Investigative Ophthalmology & Visual Science, 65(6), 10.

Chapter 6 presents machine learning models for analysis of the choroid in optical coherence tomography images. This work was undertaken with my dear colleague Jamie Burke and two papers have been published that we are "joint first authors" on, in alternating order. One (Burke et al., 2023a) in ARVO TVST:

> Burke, J.*, Engelmann, J.*, Hamid, C., Reid-Schachter, M., ..., Storkey, A., ... Bernabeu, M. O., & MacCormick, I. J. (2023). An Open-Source Deep Learning Algorithm for Efficient and Fully Automatic Analysis of the Choroid in Optical Coherence Tomography. Translational Vision Science & Technology, 12(11), 27.

And another (Engelmann et al., 2024a) in ARVO IOVS:

> Engelmann, J.*, Burke, J.*, Hamid, C., Reid-Schachter, M., ..., Storkey, A., ... Bernabeu, M. O., & MacCormick, I. J. (2024). Choroidalyzer: An Open-Source, End-to-End Pipeline for Choroidal Analysis in Optical Coherence Tomography. Investigative Ophthalmology & Visual Science, 65(6), 6.

Chapter 7 presents a machine learning model for assessing the quality of colour fundus images called QuickQual. This work (Engelmann et al., 2023a) has been published in the proceedings of the MICCAI workshop on Ophthalmic Medical Image Analysis 2023 as:

> Engelmann, J., Storkey, A., & Bernabeu, M.O. (2023). QuickQual: Lightweight, Convenient Retinal Image Quality Scoring with Off-the-Shelf Pretrained Models. In: Antony, B., Chen, H., Fang, H., Fu, H., Lee, C.S., Zheng, Y. (eds) Ophthalmic Medical Image Analysis. OMIA 2023. Lecture Notes in Computer Science, vol 14096.

Each of these papers has been reproduced as permitted by either the license it was published under or the publisher's policy for author re-use.

1.3 What's not in this thesis

This thesis makes heavy use of previously published material, as permitted by university regulations. Thus it differs in format and style from traditional theses that do not include such material, yet I believe that it demonstrates each of the qualities to be examined equally well.

Each paper contains an introduction that surveys the relevant literature and identifies a gap to be addressed, as well as a conclusion that critically reflects on the work. In compliance with university regulations, this thesis further bookends each paper by additional introductions and conclusions. Thus, I believe that my adequate knowledge of the field of study and relevant literature as well as my ability to exercise critical judgement of the work of myself and of others is sufficiently demonstrated and a dedicated literature review chapter would be superfluous.

Likewise, my capability of pursuing original research that makes a contribution to the field is embodied by the work presented here. Each paper contains a methodology section that explains the relevant methods, with Chapter 1 and Chapter 2 in particular showcasing my ability to express myself using technical language. My ability to apply machine learning methods is evidenced by their successful application. Thus, a traditional methods chapter would be similarly superfluous, in my opinion.

Finally, I think that my ability to present results in a critical and scholarly way is demonstrated by the papers themselves. Them having been accepted in reputable, peer-review venues is of course no guarantor of quality but offers some support to my belief. As explained in the previous section, the remaining text of this thesis is held in a somewhat more direct and open style, which is a personal preference of mine and more reflective scientific discourse outside of papers. My hope is that this allows the reader to assess my scientific abilities in more depth than if this entire thesis was held in the same style as the papers.

Not everything I worked on during my time as a PhD student made its way into this document. First, and most importantly, I have held two posts as a research assistant and none of the work undertaken for those is included in my thesis. Second, I did not include work that I co-authored but did not lead (Tabuchi et al., 2024; Villaplana-Velasco et al., 2023). Third, work that is already written up but not yet published in a

peer-reviewed venue has not been included either (Burke et al., 2024; Engelmann and Bernabeu, 2024). Finally, as is the case with most researchers, I have a long list of projects in various stages of completion, some of which will never see the light of day, some of which will hopefully appear as papers in the future.

Chapter 2

Machine learning for retinal disease detection in ultra-widefield fundus images

2.1 Introduction

Loss of vision due to retinal disease severely reduces people's quality of life (Brown, 1999). In a recent cross-sectional survey of UK adults, sight was the most valued sense (Enoch et al., 2019) to the point that people on average would prefer 4.6 years of perfect health over 10 years of life being completely blind. Sight loss also has major adverse economic impact (Pezzullo et al., 2018) costing the UK billions of pounds annually in healthcare cost and lost productivity, in addition to the value of healthy life lost which itself is many billions annually.

This work originated as my MSc by Research thesis project of my PhD programme. When I started working on this project, the same dataset had already been used by a number of studies (Masumoto et al., 2018a,1,1; Matsuba et al., 2019; Nagasato et al., 2018,1; Nagasawa et al., 2019,1; Ohsugi et al., 2017; Tabuchi et al., 2018) that used deep learning for disease detection. Thus, I was initially uncertain whether there would be anything for me to do that would be both meaningful and novel, especially as this was my first project in retinal image analysis. However, I noticed that those previous works had some limitations that could be addressed, primarily relating to how the problem was framed and the dataset filtered. Each of these studies had focused on binary classification of a single retinal disease and then only selected images of retinas showing this specific disease or no disease at all, while excluding all images that showed other diseases. That is not a particularly realistic setup: even in an age-related macular degeneration clinic, for example, patients might occasionally

present with other retinal diseases. A model that has not be trained on data relating to any other conditions might then classify unknown disease as age-related macular degeneration (after all, it is not healthy) or as healthy (after all, it is not age-related macular degeneration) neither of which is desirable. A secondary limitation of previous work was that poor quality images, difficult to judge cases, and images with multiple diseases were removed. However, each of these challenges are clinical realities that a model ideally should be robust to. In the case of image quality, some images are indeed too poor to be properly assessed, but including them in the dataset, especially for evaluation, will give a more pessimistic estimate of performance. Whereas excluding poor quality images from analysis runs the risk of giving an overly optimistic estimate instead if the bar for quality is set too high.

We address these limitations by developing a disease detection model that could detect multiple key retinal diseases (glaucoma, diabetic retinopathy, retinal vein occlusion, retinal detachment, age-related macular degeneration, macula hole, and retinitis pigmentosa) while also providing a prediction for whether the eye shows any disease at all. This additional prediction helps in cases where the model is confident that the image shows disease but is not very confident in which specific disease it is. It might also allow the model to flag up diseases not present in the training data. These diseases were selected in part because those were the labels provided by the dataset. However, they in turn were provided because they are key retinal diseases. For example, diabetic retinopathy, age-related macular degeneration, and glaucoma are very common sight-threatening conditions. Retinal detachments and macular holes are somewhat less common but can require very urgent interventions. The dataset also had a very small number of images labelled as artery occlusion. However, this number was so small that we could not have properly distributed them between our training, validation and internal testing sets. Instead, we held out these images entirely for our internal testing set to see how the model responds to the presence of a new disease that it did not encounter during training, which we analyse in one of the appendices.

In machine learning, for classification tasks with multiple labels, a common approach is to frame it as multi-class classification, where each input is assumed to belong to exactly one class. There, the output of the model is put through a softmax activation function which normalises all the outputs to sum to 1. However, in our case, an image could show more than one disease, thus this constraint is undesirable. Instead we

frame the problem as multi-label classification and apply an element-wise sigmoid activation function (equivalent to a logistic linkage function) such that each output individually is constrained to $(0, 1)$. This is straight-forward conceptually and in implementation, but less well-known than binary or multi-class classification.

When evaluating the model, we keep all of the data including difficult cases and low quality images, as we prefer to have a potentially pessimistic performance estimate over a potentially optimistic one. In addition to quantitative evaluation, we manually examine cases of confident errors to understand where the model might perform poorly. We also sketch out two archetypical use cases and look at performance at two corresponding potential operating points on the receiver operating characteristic curve. Furthermore, we investigate which regions of the images are important for model performance, both to validate that the model appears to work in a sensible way and to understand whether the increased field of view is beneficial for machine learning-based disease detection. The latter question is particularly interesting as ultra-widefield fundus cameras are substantially more expensive than standard field ones.

2.2 Paper

Please note, as mentioned in section 1.2 of this thesis, this work originated as my thesis for the MSc by Research stage of my PhD programme, but underwent substantial changes in style and writing, as well as refinement of the analysis prior to submission and publication in Nature Machine Intelligence. I worked on these changes during the PhD stage of my programme. Thus, I am of the opinion that this work is relevant to this PhD thesis.

nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00566-5

Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning

Received: 17 December 2021

Accepted: 14 October 2022

Published online: 8 December 2022

Check for updates

Justin Engelmann[1,2] ✉, Alice D. McTrusty[3,4], Ian J. C. MacCormick[5], Emma Pead[3], Amos Storkey[2,7] & Miguel O. Bernabeu[1,6,7] ✉

Ultra-widefield (UWF) imaging is a promising modality that captures a larger retinal field of view compared with traditional fundus photography. Previous studies have shown that deep learning models are effective for detecting retinal disease in UWF images, but primarily considered individual diseases under less-than-realistic conditions (excluding images with other diseases, artefacts, comorbidities or borderline cases; and balancing healthy and diseased images) and did not systematically investigate which regions of the UWF images are relevant for disease detection. Here we first improve on the state of the field by proposing a deep learning model that can recognize multiple retinal diseases under more realistic conditions than what has previously been considered. We then use global explainability methods to identify which regions of the UWF images the model generally attends to. Our model performs very well, separating between healthy and diseased retinas with an area under the receiver operating characteristic curve (AUC) of 0.9196 (±0.0001) on an internal test set, and an AUC of 0.9848 (±0.0004) on a challenging, external test set. When diagnosing specific diseases, the model attends to regions where we would expect those diseases to occur. We further identify the posterior pole as the most important region in a purely data-driven fashion. Surprisingly, 10% of the image around the posterior pole is sufficient for achieving comparable performance across all labels to having the full images available.

Retinal diseases are a key public health burden. The associated loss of vision reduces the quality of life of affected patients[1] and has major economic impact[2]. Age, lifestyle factors and some diseases (for example, diabetes) are key risk factors for retinal diseases such as age-related macular degeneration and diabetic retinopathy. This burden is thus set to increase in ageing societies and in those that are rapidly urbanizing and experience rising incidence of non-communicable diseases. Retinal diseases are detected and diagnosed through eye examinations, which

[1]Centre for Medical Informatics, Usher Institute, The University of Edinburgh, Edinburgh, Scotland. [2]Institute for Adaptive and Neural Computation, School of Informatics, The University of Edinburgh, Edinburgh, Scotland. [3]Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, Scotland. [4]NHS Education for Scotland, Edinburgh, Scotland. [5]Centre for Inflammation Research, The University of Edinburgh, Edinburgh, Scotland. [6]The Bayes Centre, The University of Edinburgh, Edinburgh, Scotland. [7]These authors jointly supervised this work: Amos Storkey, Miguel O. Bernabeu. ✉e-mail: justin.engelmann@ed.ac.uk; miguel.bernabeu@ed.ac.uk

Table 1 | Overview of the TOP dataset and the three subsets

	TOP dataset	Train set	Validation set	Test set
Patient classification				
Healthy	2,322 (43.2%)	1,625 (43.2%)	348 (43.2%)	349 (43.2%)
Gla	943 (17.5%)	660 (17.5%)	141 (17.5%)	142 (17.6%)
DR	682 (12.7%)	477 (12.7%)	102 (12.7%)	103 (12.8%)
RVO	438 (8.1%)	307 (8.2%)	65 (8.1%)	66 (8.2%)
RD	417 (7.8%)	292 (7.8%)	63 (7.8%)	62 (7.7%)
AMD	285 (5.3%)	200 (5.3%)	43 (5.3%)	42 (5.2%)
MH	179 (3.3%)	125 (3.3%)	27 (3.3%)	27 (3.3%)
RP	110 (2.0%)	77 (2.0%)	17 (2.1%)	16 (2.0%)
Total patients	5,376	3,763	806	807
Total eyes	8,570	6,002	1,279	1,289
Total images	13,026	9,121	1,911	1,994
Patient age (mean±s.d.)	65.83±12.99	65.72±13.04	65.89±12.84	66.29±12.87
Female patients	2,669 (50%)	1,843 (49%)	392 (49%)	434 (54%)

We report the number and share of patients by stratification value ('Details of data split with patient-level stratification' in Methods), the number of patients and images, as well as mean age and sex ratio for each set.

often include fundus imaging. Traditional colour fundus photography (CFP) typically images 30–60° of the retina, but new imaging devices allow for ultra-widefield (UWF) retinal images with a field of view of 100–200° (refs. [3,4]). Thus, UWF could capture retinal pathology that would be missed with CFP. This could enable more accurate screening and earlier detection. Sophisticated retinal imaging hardware is increasingly available in community optometry practice. Diagnosis and support systems are urgently needed to help practitioners make clinical decisions using these new imaging modalities, particularly for UWF retinal images that fewer clinicians are experienced with.

Fortunately, recent advances in deep learning (DL) make it feasible to automatically process images with similar performance to human experts on many tasks. Previous work has shown that DL models are effective at detecting disease in UWF images but the evaluation approaches employed in literature so far have been criticized as unrealistic by clinical practitioners[5]. For example, models were typically trained to detect only a single specific disease and then evaluated on a dataset that contained equal numbers of images showing that disease and healthy controls, but no other diseases[6–15]. Recent work proposed a model for three diseases, but also excluded other diseases and artificially balanced the data[16]. Images with artefacts or multiple diseases were also excluded[6–16]. We improve on this approach by proposing a model that can detect seven different retinal diseases and under more realistic conditions than what has previously been considered. In particular, we consider several retinal diseases without excluding images with multiple diseases in the same image, artefacts or borderline cases. We further do not artificially balance the data by discarding healthy cases.

Thus far, there has been no systematic investigation of which UWF image regions are relevant for DL model-based detection of retinal disease. Previous work compared the performance of DL models for detecting a single disease (glaucoma) using either full or optic-disc-cropped UWF images[9]—but in this case, the optic disc was selected based on domain knowledge. Similar work has been undertaken for CFP, but again the regions of interest were defined using domain knowledge[17]. We instead investigate which regions the DL model attends to for seven different diseases through a data-driven analysis. This serves both to understand how DL models use UWF images and to validate that the DL model works as intended rather than relying on undesirable shortcut artefacts. We further investigate how model performance degrades when removing either the least or the most important regions.

We find that the proposed DL model performs very well on an internal test set even when this is not artificially balanced and includes images with artefacts, borderline cases and comorbidities. Our model also outperforms an ensemble of binary classifiers trained on balanced data for individual diseases, which is the prevailing approach in the literature. Evaluation on a challenging external test set that includes images with different preprocessing and images taken with a variety of UWF imaging device models also evidences very high real-world performance. For individual diseases, the model attends to regions that are qualitatively consistent with domain knowledge, indicating that the model works as intended. Interestingly, we find that all seven diseases are detectable in the posterior pole and arrive at this conclusion in a purely data-driven way.

Results

Study data characteristics and data split

The Tsukazaki Optos Public (TOP) dataset consists of 13,047 UWF retina images of 8,588 eyes belonging to 5,389 patients that were taken between 11 October 2011 and 6 September 2018 at Tsukazaki Hospital in Himeji, Japan. All images were taken with an Optos 200Tx (Dunfermline, Scotland) UWF scanning laser ophthalmoscopy imaging device. The data were released by Dr Hiroki Masumoto for research use and this was approved by the Ethics Committee of Tsukazaki Hospital. Each image has binary labels for eight retinal diseases: artery occlusion (AO), age-related macular degeneration (AMD), diabetic retinopathy (DR), glaucoma (Gla), macular hole (MH), retinal detachment (RD), retinitis pigmentosa (RP) and retinal vein occlusion (RVO). For each image, the dataset also contains a unique patient ID, age in years, sex, whether the image is of a left or a right eye, and the binary diabetes status of the patient. 50% of the patients are female, and 50% of the images are of left eyes. The average patient age is 65.83 yr with a standard deviation of 12.99 yr. Table 1 shows the numbers of patients per disease.

Apart from the 21 images showing AO, we did not exclude any images from the study. These images were excluded from the main analysis as 21 is too few to both train and meaningfully evaluate a model on that label. Instead, those images were used as an additional, held-out test set to assess whether our model generalizes to unseen diseases (Supplementary Section 4). Common reasons for excluding images in the literature are quality issues (for example, reflection or eyelash artefacts, poor contrast), difficult-to-recognize pathology (for example, borderline cases), multiple diseases or discarding of healthy controls

https://doi.org/10.1038/s42256-022-00566-5

Table 2 | Test set performance of baselines and final model for each label

(a) Test set performance using AUC as metric	Diseased	DR	Gla	RD	RVO	AMD	RP	MH
Logistic regression with Age+Sex	0.6017±0.0003	0.6008±0.0003	0.5191±0.0004	0.7555±0.0005	0.4946±0.0007	0.7989±0.0006	0.6625±0.0010	0.5630±0.0010
Ensemble of experts (binary DL models+balanced data)	0.8346[a]±0.0002	0.8405±0.0003	0.9140±0.0002	0.9281±0.0005	0.8922±0.0005	0.7068±0.0008	**0.9490**±0.0003	0.6300±0.0010
Ours (single multi-label DL model+realistic data)	**0.9196**±0.0001	**0.9153**±0.0002	**0.9452**±0.0002	**0.9788**±0.0002	**0.9407**±0.0003	**0.9503**±0.0003	0.9408±0.0005	**0.7939**±0.0009

(b) Test set performance using Brier score as metric	Diseased	DR	Gla	RD	RVO	AMD	RP	MH
Logistic regression with Age+Sex	0.2451±0.0001	0.1488±0.0001	0.1634±0.0001	0.0491±0.0001	0.0555±0.0001	0.0388±0.0001	0.0244±0.0001	**0.0210**±0.0001
Ensemble of experts (multiple binary DL models+balanced data)	0.1850[a]±0.0001	0.1474±0.0001	0.1213±0.0001	0.0820±0.0001	0.1275±0.0001	0.2090±0.0001	0.1038±0.0001	0.2421±0.0001
Ours (single multi-label DL model+realistic data)	**0.1146**±0.0001	**0.0741**±0.0001	**0.0651**±0.0001	**0.0146**±0.0000	**0.0280**±0.0001	**0.0248**±0.0001	**0.0079**±0.0000	0.0223±0.0001

Reported values are mean±s.e., obtained by bootstrapping the data 1,000 times. Images are weighted such that each eye has a total weight of 1, even if a specific eye was imaged multiple times. AUC assesses how well a model can separate positive and negative samples for a given label. Higher is better; best values are in bold. Brier score is sensitive to how well a model's predicted probabilities are calibrated. Lower is better; best values are in bold. [a]Using maximum of individual predictions ('Benchmark models' in Methods).

to balance the labels[6–16]. We decided against excluding images for any of those reasons as poor-quality images, borderline cases, comorbidities and label imbalance are clinical realities and—as recently highlighted by clinical practitioners[5]—DL models should be robust to those challenges. Furthermore, setting the bar for image quality high runs the risk of an overly optimistic estimate of DL model performance, whereas our approach is more likely to provide an estimate that is too pessimistic. We consider this preferable when developing methods intended for clinical applications.

We split the data into three sets—training, validation and test—containing 70%, 15% and 15% of the patients, respectively. Including a validation set is an important methodological detail that has been omitted by some previous work on the TOP dataset[6–15]. Developing DL models is typically an iterative process and the validation set guides the process while keeping the test set unseen. Not using a validation set in machine learning is comparable to not controlling for multiple comparisons in statistics. We chose our final model on the validation set before making any test set evaluations. Note that in the machine learning community, the final held-out set used to estimate model performance is called the 'test set', whereas the set used to provide feedback for the modelling process is called the 'validation set'. This clashes with terminology used in the medical community, where the evaluation on the test set is typically referred to as 'internal validation'. Note that the validation set (in the machine learning sense) is not used for internal validation (in the medical sense). Furthermore, we split the data on the patient level rather than image level to prevent images of the same eye appearing in different sets.

To ensure that the distributions of diseases are similar across all three sets, we split stratified by patient disease status (see 'Details of data split with patient-level stratification' in Methods).

Our model achieves state-of-the-art results

The DL model outputs a probability for each of the seven retinal diseases as well as a probability for the input retina being diseased generally. To allow our model to deal with images that show multiple diseases, we frame the problem as multi-label rather than multi-class classification. This allows the model to predict more than one disease with high confidence, if appropriate. Our approach is described in detail in 'Model and problem framing' in Methods. We evaluated our model on the unseen internal test set. Table 2a shows the area under the receiver operating characteristic curve (AUC) for each label. We weigh images such that each eye contributes equally to the metrics, regardless of how many times it was imaged. Our model can discriminate between healthy and diseased retinas with an AUC of 0.9206 and achieves excellent discrimination for six diseases (AUCs 0.9125–0.9753) and good discrimination (AUC = 0.7987) for the seventh (MH). MH is the disease with the fewest images available and thus our model had the least examples to learn from. Furthermore, diagnosis of MH would commonly be confirmed using different types of imaging such as optical coherence tomography and can be difficult to recognize from images alone[18].

We also consider two baselines ('Benchmark models' in Methods). First, a model using only the available patient information (that is, age and sex) as variables but no image information. We include this as a simple baseline to ensure that the DL model does not only infer age and sex from the image to make its predictions. We considered random forest, *k*-nearest neighbours and logistic regression as classification algorithms and found that logistic regression performed best on the validation set. Second, an 'ensemble of experts' of binary models that are the same as our proposed model except that they each have a single output and were trained on balanced data containing only controls and a specific disease. This is emulates approaches that have been used in previous work[6–15]. We find that the DL models clearly outperform the Age + Sex baseline on all labels, except on AMD where the ensemble of experts performs the worst. The ensemble of experts achieves a slightly higher (+0.0052) but similar AUC on RP, but our proposed model substantially outperforms it on all other labels. We also assessed calibration through the Brier score (Table 2b) and find that the ensemble of experts is poorly calibrated compared with our proposed model, with Brier scores on average being 5.6-times higher. This is a result of being trained on artificially balanced data.

In addition to technical measures of model performance, we also evaluated our model for two archetypical use cases to assess whether the model would be useful in practical applications[5].

First, a use case where false positives are costly and we thus want a conservative decision threshold $\hat{p}^t$ for flagging an image up as diseased, for example, an early screening application at a high-street optician where a false positive would lead to an unnecessary referral. Second, a use case where false positives are less costly and we want a less conservative decision threshold, for example, a clinical decision support system at a specialist clinic where false positives can be quickly dismissed and the focus is on reducing the chance that something is overlooked. In a concrete application, any point on the Receiver

https://doi.org/10.1038/s42256-022-00566-5

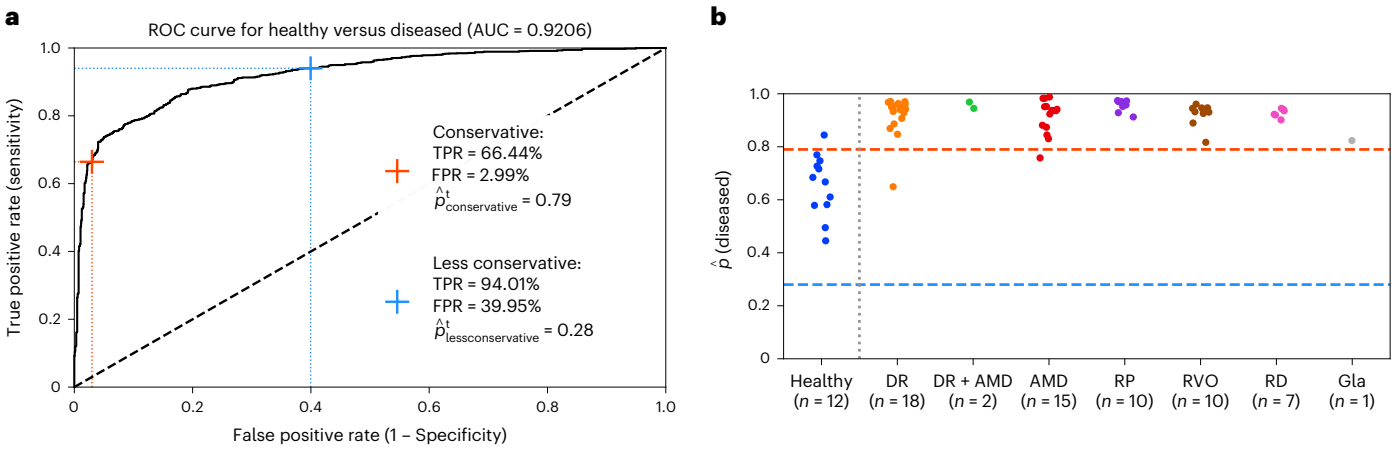


Fig. 1 | Evaluation of our model on the test set and the external validation set. **a**, ROC curve of our model predicting healthy versus diseased on the test set, weighted such that each eye contributes equally. Markers indicate where on the ROC curve we end up given the constraint of the respective use case. The dashed line is the identity line equivalent to random guessing. **b**, External validation results showing the predicted probability of being diseased $\hat{p}$(diseased) stratified by ground-truth label. The red and blue horizontal line plot indicates the conservative threshold $\hat{p}^{t}_{conservative} = 0.79$ and and less conservative threshold $\hat{p}^{t}_{lessconservative} = 0.28$, respectively.

Operator Characteristic (ROC) curve can be chosen, and we could implement a complex, multi-level decision process. For example, we could use a traffic-light-like system where a yellow alarm is raised for samples where $\hat{p}$ exceeds a less conservative threshold and a red alarm if it exceeds a conservative threshold. In an early screening context, the yellow alarm might then translate into simply scheduling the next routine scan sooner and the red alarm means that the patient will be referred to an ophthalmologist. In any case, the clinician ultimately makes the diagnosis and decides on the next steps taking into account the available information. This includes their own assessment of the patient and the patient's health history and symptoms. A DL-based clinical decision support system would be an additional input to the clinician's decision-making. Developing such a tool from the DL model requires additional considerations such as the available hardware to run the model on, what information to display to the clinician and how to train them in interpreting the tool's output[19]. Considering these aspects in detail is beyond the scope of the present work, but we aim to investigate whether the performance at particular decision thresholds would be potentially practically useful.

For the present work, however, our goal is to use the two archetypical use cases—a conservative and a less conservative case—to evaluate whether our model's performance at concrete decision thresholds would be potentially clinically useful. For the conservative and less conservative cases, we take a maximal false positive rate (FPR) of 3% and 40% (equivalent to minimum specificities of 97% and 60%), respectively, as our constraints. We then choose the optimal threshold that maximizes the true positive rate (TPR) given these constraints. Figure 1a shows the ROC curve obtained by our model, and reports the resulting TPRs, FPRs and decision thresholds $\hat{p}^{t}$. In the conservative case, the model could detect about two-thirds of patients with diseased retinas while only incorrectly classifying 3% of healthy retinas as diseased. In the less conservative case, where we prioritize a high TPR over a low FPR, the model could detect 94% of diseased retinas while incorrectly flagging up 40% of healthy retinas as diseased. We thus conclude that our model would potentially be useful in clinical applications and performs very well overall, especially considering the more challenging test set compared with what has previously been considered.

Our model performs very well on a difficult external dataset

Following the example of recent work[16], we assembled an external dataset of 75 UWF images. For a detailed description, see Supplementary Section 3. Although small, this external dataset is quite challenging in a number of ways. First, most images are taken with different models of UWF imaging devices (for example, Optos Daytona or California) that produce qualitatively different images. Second, many of the images are cropped (and thus also have a different scale), projected, have a different aspect ratio or even contain watermarks. We did not correct for these factors through preprocessing. Third, these images are likely to not be of Japanese patients and thus present a different patient population.

Figure 1b shows the predicted probability of being diseased stratified by ground-truth label. Despite these challenging conditions, our model achieved near perfect discrimination between healthy and diseased retinas (AUC = 0.9848 ± 0.0004). Using the more conservative early screening threshold, our model would have had one false positive, two false negatives and correctly classified all remaining 72 images. However, we also find that calibration of our model suffered under these conditions, and using the less conservative threshold, we would have recommended all images for further investigation. This might be due to all images looking potentially anomalous to our model as they are quite different from the images it was trained on. The model also identified the correct disease(s) in most cases, achieving very high label-wise AUCs for all included diseases (DR, 0.9636 ± 0.0007; RP, 1.0000 ± 0.0000; RVO, 0.9770 ± 0.0007; RD, 1.0000 ± 0.0000; Gla, 0.9599 ± 0.0008) apart from AMD where it achieved an AUC of 0.8033 ± 0.0021, which is good but noticeably worse than for the other labels. This might indicate that differences in imaging devices and data preprocessing have a larger impact on recognizing AMD as such but it could also be due to differences in diagnostic thresholds. Overall, this is very encouraging performance on an external dataset that differs substantially from our training data.

Model attends to regions consistent with domain knowledge

We used explainable artificial intelligence (AI) techniques to understand how the model makes its predictions. First, we use gradient-weighted class activation mapping (GradCAM)[20] to generate attention maps for a given input image and target disease. These highlight which regions of the image the model considers evidence for the respective disease. We find that the model generally pays attention to regions of pathology even in the presence of distractions such as reflection artefacts (Fig. 2a) but sometimes fails in the presence of these (Fig. 2b). Overall, these maps could aid clinicians in practice to understand the

https://doi.org/10.1038/s42256-022-00566-5

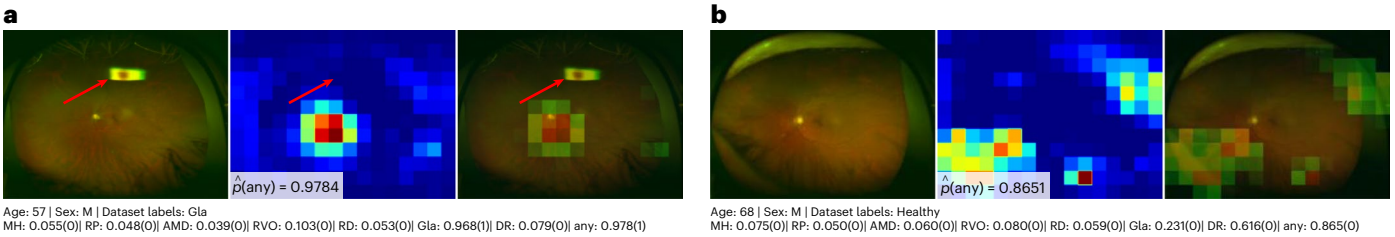


Fig. 2 | Examples of the model-attention heat maps generated by GradCAM using the general 'diseased' label as target concept. Left: Original image. Middle: Heat map. Right: Heat map imposed on the image. The first line of text below the image indicates the patient's age, sex and disease status according to the dataset labels. The second line of text indicates the model predictions for each label and the dataset labels in brackets. **a**, An example of the model successfully ignoring a reflection artefact (highlighted by a red arrow) and instead attending to the optic disc to correctly detect that the retina is diseased and that the disease in particular is Gla ($\hat{p}(\text{Gla}) = 0.968$). **b**, An example of the model focusing on artefacts, particularly eyelash artefacts, in the periphery and falsely predicting a high probability of being diseased. The predicted probability of DR is highest ($\hat{p}(\text{DR}) = 0.616$), which might be due to the eyelash artefacts being visually similar to haemorrhages or microaneurysms. There is also a small bright spot southwest of the optic disc, which receives attention and might be interpreted as an exudate.

model's predictions, to draw their attention to regions of interest on an individual patient basis or to identify whether an artefact might be adversely affecting the model's prediction.

While individual image-wise attention maps are useful for understanding the model, only a small number of images can be examined in detail and reproduced at once, which constitutes a barrier to auditing DL models thoroughly and objectively. To examine the model more systematically and to understand which regions of the UWF images are useful to detect specific diseases, we generate label-wise global attention maps through predicted probability-weighted aggregation of image-wise attention maps[21]. These maps summarize the image-wise attention maps across the entire test set. Figure 3a shows the resulting maps for the seven diseases and the general diseased label. These maps are consistent with domain knowledge. For example, the model focuses on the posterior pole for both Gla and AMD but for Gla this is concentrated on the optic disc side whereas for AMD this is concentrated on the macula. For RD, there is attention across the entire retina but concentrated on the temporal side where rhegmatogenous RD tends to occur most often[22]. For DR, there also is attention across the retina but clearly concentrated around the optic disc, presumably due to cases of proliferate DR with neovascularization of the optic disc. Finally, although the map for MH is concentrated on the macula, it is also the noisiest map. This is due MH being the most difficult label for our model and due to the low number of images we average over as MH was the rarest disease in terms of images. To summarize, these maps generally match domain knowledge. This indicates that the model detects pathology for specific diseases in regions where we would expect pathology to occur.

Peripheral regions are not needed for high performance

We combined all label-wise global attention maps into a single global attention map (Fig. 3b)[21]. This map ranks each position of the input images in terms of their overall importance to the model and identifies the posterior pole as the most important region. This agrees well with domain knowledge, but we note that this was identified in a purely data-driven way. The global attention map also correctly identifies the corners that never show the retina as least important.

We use progressive erasure plus progressive restoration (PEPPR)[21] to validate that the global attention map is faithful to how the model makes its predictions and to investigate how the model's performance degrades when removing either the most or the least important regions. Figure 4 shows the AUC for each label for the test set with different parts of the image removed. When progressively erasing the least important image regions according to the global attention map, there is no significant drop in AUC even when 90% of the image is removed. This indicates that this map does correctly reflect which regions are most important to the model. However, it is surprising that having only the most important 10% of the UWF images is sufficient to obtain performance comparable to having the full images available. This suggests that all seven diseases have presentation in the posterior pole, which might be in part a reflection of the disease stages common in the TOP dataset. When progressively restoring the least important regions, starting with a blank image, we observe a near monotonic increase in AUCs with performance only peaking for all labels once the full image has been resorted. For diseases that primarily affect the optic disc and fovea (Gla, MH, RVO and AMD), restoring the final most important 10% containing the posterior pole leads to a large increase in AUC (approximately >0.1), whereas for diseases that cause more peripheral pathology (RD, RP and DR) there is a less substantial increase (approximately <0.03). In summary, we find that the global attention map faithfully reflects how the model works. Surprisingly, the 10% of the image containing the posterior pole is sufficient to achieve high performance.

Discussion

We find that DL is very effective for detecting disease in UWF images even when evaluating the model under more realistic conditions than what has previously been considered. This strong performance was also confirmed on a challenging external test set, including images taken with different cameras, different aspect ratios or even watermarks. In practice, we would not recommend applying a model to images that are from different imaging devices or preprocessed in ways the model has not encountered during training. It is not unexpected that calibration would suffer due to this, but the excellent discrimination in terms of AUC is very promising. We also note that we decided on the decision thresholds before the external dataset was collected, so it is remarkable that our model would have made only 3 mistakes out of 75 images using the conservative threshold. Thus, our DL model shows promise for applications such as early screening and clinical decision support. Using more realistic conditions (no artificial balancing of data, multiple diseases, comorbidities, no exclusion of difficult cases) for evaluation also resulted in lower AUCs than other published models[6–15]. To ensure that this is due to the test set being more challenging rather than our approach being inefficient, we compared our model against an ensemble of identical models that were trained as binary classifiers on individual diseases using artificially balanced data as is commonly done in the literature. We find that our approach clearly outperforms this approach both in terms of separation (AUC) and especially calibration (Brier score). For one previous study on the TOP dataset focused on RP[15], the exact subset of the TOP dataset used after excluding images is available. On this subset, a simplified version of our model achieved perfect separation (AUC = 1) with very high confidence in the correct

https://doi.org/10.1038/s42256-022-00566-5

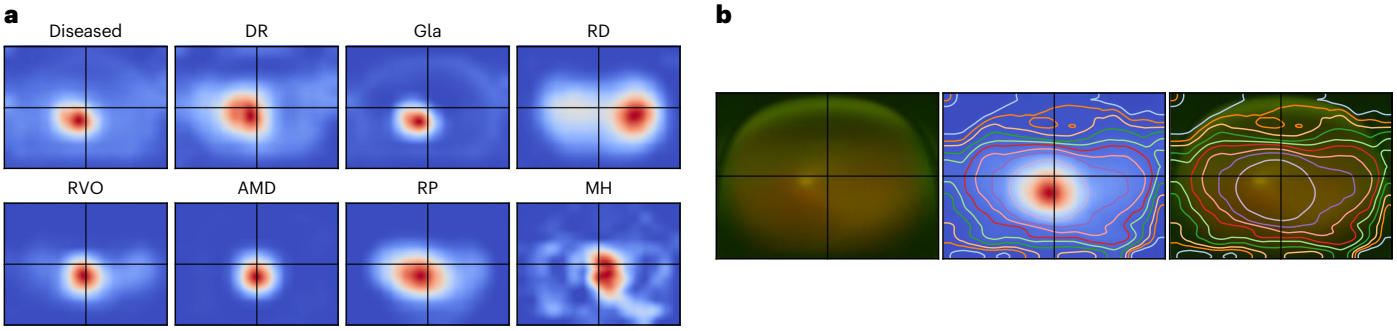


Fig. 3 | Global attention maps of our model. a, Label-wise global attention maps. Calculated on all images from the test set. Right eyes are flipped horizontally such that the temporal side is on the right for all images. Blue indicates regions of low attention and red indicates regions of high attention. The black crosshairs are added to allow for easier comparison of relative positions between subplots and are centred on the image centre, which approximately corresponds to the macula. **b**, Left: average of all test set images. Right eyes are flipped horizontally such that the temporal side is on the right for all images. Middle: global attention map. Blue indicates regions of low attention and red indicates regions of high attention. Contour lines indicate the most important regions in quantile steps of 10%. Right: the average of all test set images with the same contour lines. The black crosshairs are added to allow for easier comparison of relative positions between subplots and are centred on the image centre, which approximately corresponds to the macula.

labels (Supplementary Section 5), which has also been noted in the literature[16]. Under more realistic and challenging conditions, however, our more complex model achieved an AUC of 0.9438 on the RP label, which is still high, but not perfect, performance. Thus, we recommend using training and validation data that are as realistic as possible given the constraint of the available data, which has also been raised by clinicians[5]. While this might yield lower performance numbers, those numbers are also more realistic and models trained on realistic data will fare better under realistic conditions.

We find that the model focuses on regions to diagnose specific diseases that are consistent with where we would expect pathology to occur. This concordance between domain knowledge and data-driven investigation implies that our model works in a desirable fashion and is unlikely to rely on 'shortcut' artefacts, which can be a problem in medical imaging[23]. It also validates current domain knowledge, for instance, we identify the posterior pole as the most important region purely from the data. Thus, this kind of approach might also be useful for knowledge discovery in the future.

Surprisingly, the posterior pole region itself is sufficient to obtain high performance. This raises questions around whether and how all seven diseases we considered have presentation there. It also raises the question of how much benefit UWF imaging has over traditional CFP. It is possible that cases in the TOP dataset might skew towards being severe or progressed, which can be detected in the posterior pole, whereas early signs might exclusively occur in the periphery. However, it has been noted that the TOP dataset contains both obvious and subtle cases of pathology at least for RD, RP and RVO[16]. It is also possible that the DL model can spot subtle signs of early pathology in the posterior pole that have not been previously noted. DL has previously been shown to be able to extract information from fundus images that was not known to be present in those images, such as sex and cardiovascular disease risk[24,25]. We also want to stress that AI-based diagnosis is far from the only use case of fundus imaging, and the additional retinal coverage of UWF images might help clinicians in making their own diagnosis, judging disease severity and choosing appropriate interventions.

Our work has several limitations, primarily due to constraints stemming from the available data. The TOP dataset is a very valuable resource, but unfortunately, few details about how it was curated (inclusion and exclusion criteria), whether each image was graded by multiple graders, or diagnostic thresholds for assigning the binary disease labels (for example, for AMD and DR) are available. These are key limitations of this dataset. In SupplementarySection 1, we present additional exploratory data analysis where we describe signs of possible selection bias: half of the patients are female despite most patients being of advanced age where we would expect more females, older patients tend to be healthier, and patients with diabetes appear at lower risk for non-DR retinal disease. To the best of our knowledge, our analysis constitutes the most comprehensive critical evaluation of the TOP dataset so far. Despite these limitations, it is still a very valuable resource to the research community and the only publicly available large-scale UWF dataset[26]. Previous work on the TOP dataset by ophthalmologists noted that the dataset contains both obvious and subtle cases, at least for the diseases that were considered there (RD, RP and RVO)[16]. However, due to being a specialist clinic, patients at Tsukazaki Hospital might skew towards severe cases. Presumably, most or all patients are Japanese and thus the population is relatively homogeneous in terms of genetics, culture-induced lifestyle factors and access to healthcare.

Future work could investigate in more detail how and where diseases present themselves in the posterior pole and whether the detection of less severe cases benefits more from having the periphery available. The latter could be achieved using the TOP dataset through manual curation of additional labels for disease severity by domain experts. Access to additional large labelled UWF datasets would allow for a more comprehensive external validation of our model. We did contact authors of previous studies using private UWF datasets. Unfortunately, we have been unable to access any such datasets thus far. We encourage the community to make datasets available whenever possible. To evaluate the benefit of UWF over CFP in more detail, it would be particularly interesting to collect a dataset of CFPs and UWF images showing the same retinas at the same point in time to compare the two modalities directly. DL models designed for community screening should ideally be evaluated on datasets that closely resemble that use case, that is, with a large proportion of healthy images and where the diseased images are primarily early or mild cases. However, such UWF datasets are not currently available to the best of our knowledge.

It might be possible to improve on the performance of our proposed model, for instance, by using extensive computing resources to manually improve the training procedure, do automatic hyperparameter tuning or train multiple models for an ensemble. However, we think that the performance we obtained might be close to the saturation point for the TOP dataset. We examined the 20 most confident false positives of our model and found that at least 14 of them are probably pathological. For details, see Supplementary Section 6. This also implies that our reported performance numbers underestimate the true model skill. Beyond outright label mistakes, with binary labels

https://doi.org/10.1038/s42256-022-00566-5

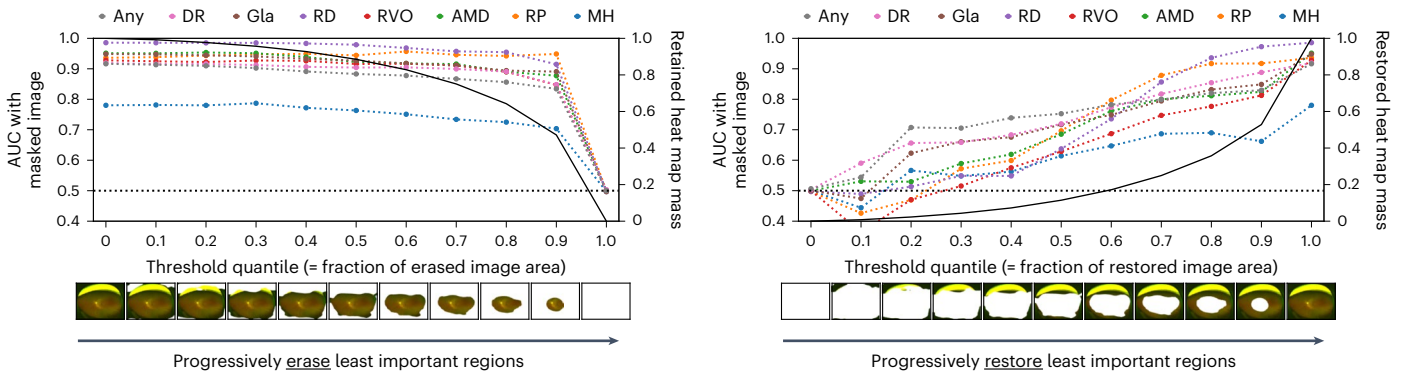


Fig. 4 | Validating the global attention maps through progressive erasure and progressive restoration. Left: progressive erasure. Right: progressive restoration. Coloured lines indicate the test set AUCs when erasing using the mask obtained with the respective threshold. The horizontal dashed black line indicates AUC = 0.5, which is equivalent to random guessing. The solid black line indicates the mass of the global average heat map at the respective threshold, using the secondary *y* axis on the right. An example image with the mask at the given threshold applied is shown below the *x* axes.

there is also conceptual ambiguity whether a borderline case is classified as pathological. Finally, improved technical performance on internal validation might not be particularly meaningful for clinical practice as the population a model encounters in practice will always be at least slightly different from the population the training data is from. This distribution shift is likely to wipe out marginal gains from extensive tuning.

The data-driven methods we used to identify globally informative regions could also be applied to CFP images, where to our knowledge previous work only defined global regions of interest using domain knowledge in advance as opposed to taking a data-driven approach[17]. Data-driven approaches have been applied to identify informative regions at the image level[27,28] but to our knowledge not globally. Furthermore, identifying informative regions in medical imaging could be combined with data augmentation methods such as mixup[29] to allow for more data-efficient model training. Finally, the PEPPR analysis here focused on explaining the proposed model and investigated which regions are important to the model at hand. Future work could repeat this with retraining the model at every step to investigate which regions of the images contain information that can be leveraged by a DL model.

We proposed a DL model and evaluated it under more challenging and realistic conditions than what has previously been considered. We find that the model is very effective at detecting diseased retinas and at identifying the correct disease(s). Thus, such models could be used in clinical practice. We further investigated which regions of the UWF image attends to and find that this matches domain knowledge. For instance, the posterior pole was identified as the most important region in a purely data-driven way. This indicates that the model works as intended. Surprisingly, using just the posterior pole region is sufficient to obtain high performance, which should be investigated further in future research.

Methods

Details of data split with patient-level stratification

We split the data at the patient level rather than the image level. This is to avoid leaking information between the sets by having images of the same eyes or same patients across different sets. An example of multiple images of the same eye in the TOP dataset is shown in Extended Data Fig. 1. If we split at the image level, images of this patient might end up in both the training set and the test set. A model that overfits ('memorizes') a patient's unique vasculature or the specific morphology of their pathology might then outperform another model that does not do this and would generalize better to unseen patients. With our approach, each patient occurs in exactly one of the three sets.

As some of the included diseases are relatively rare, we conduct a stratified split to ensure that the distribution of the diseases is as similar as possible across sets. However, as we have multiple, non-exclusive labels per image and multiple images per patient, stratification is more complex than if we had a single label per patient. Thus, we assign a stratification label to each patient according to the following three rules. First, if the patient has any disease across all their images, we take the disease that occurs most frequently as their stratification label. Second, if there is a tie (that is, two or more diseases occur equally often in a given patient), we take the rarest of the diseases as their stratification label. Third, if and only if a patient has no disease in any of their images, we assign 'healthy' as their stratification label. This gives us a single label with eight possible values, the seven diseases and 'healthy', per patient and we can then stratify on that.

Model and problem framing

To develop a clinically useful model that can deal with comorbidities, we frame the problem as multi-label classification as opposed to binary[6–15] or multi-class[16] classification that have been used in previous work. The DL model outputs a probability for each label in both multi-class and multi-label classification. However, in the multi-class case, a softmax output activation is used, which means that the model always outputs a total probability mass of 1 across all outputs, and thus an output for 'healthy control' must be included. In the multi-label case, however, an element-wise sigmoid activation is applied to each output, which allows the model to allocate a probability mass between 0 and 1 per label, and a total probability mass between 0 and n_{labels} across all labels. The sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ maps the raw model outputs $x \in \mathbb{R}$, called logits, to a predicted probability $\hat{p} \in (0,1)$.

Concretely, this allows the model to detect more than one disease in a given image. Suppose an image shows DR and AMD, then with an element-wise sigmoid activation the model can output high predicted probabilities $\hat{p}$ for both labels, whereas with a softmax activation it can only output moderate values of $\hat{p}$ for both classes or a high $\hat{p}$ for one class and a low $\hat{p}$ for the other. Likewise, suppose the model has only medium confidence that both of these diseases are present in the image. With an element-wise sigmoid, it can then output moderate values of $\hat{p}$ for both to express this. We use one label for each of the seven retinal diseases and also include a label for being 'diseased' generally. This label allows the model to indicate that it is confident that a given image is affected by disease, whether it is confident in any disease in particular or not. Furthermore, this label is also the most clinically relevant label if we want to decide whether a patient needs to be referred to an ophthalmologist for further examination.

As our model, we use a convolutional neural network backbone that extracts a feature vector from an image and a prediction head that maps this feature vector to the eight-dimensional output to which we then apply element-wise sigmoid activations to obtain predicted probabilities for each label. We use ResNet34[30] as the convolutional backbone, which consists of 33 convolutional layers with residual connections followed by an average pooling layer and a linear output layer that we replace with our own prediction head. After the last convolutional layer, ResNet34 outputs a three-dimensional feature map, 2 spatial image dimensions and 512 channels, which the average pool converts to a flat 512-dimensional feature vector by averaging across the spatial dimensions. We use ResNet34 in its original configuration with batch normalization[31] and rectified linear units (ReLUs) as the activation function, where $\mathrm{ReLU}(x) = \max(0, x)$. We chose ResNet as it is a high-performance, efficient and well-established architecture that has recently been shown to be very competitive when using modern training techniques[32,33], and ResNet34 in particular as it was the largest ResNet variant that fitted into a graphics processing unit (GPU) memory at a reasonable batch size. We initially experimented with ResNet18 and found that moving to ResNet34 offered a very minor performance gain on the validation set. Thus, we expect that using larger models would not yield a significant performance improvement given that the dataset is small by DL standards.

Our prediction head is a feed-forward neural network with one hidden layer containing 128 hidden units and 8 outputs, one for each label. As the activation function for the hidden units in the prediction head, we use parametric rectified linear units (PReLU)[34], where $\mathrm{PReLU}(x) = \max(0, x) + \alpha\min(0, x)$, which extends ReLU by not zeroing out negative inputs completely, which improves convergence and where the negative slope parameter α is itself a learnable parameter. We learn a separate α for each unit of the hidden layer. Using a prediction head with a hidden layer rather than linear head allows our model to more easily learn appropriate correlations between labels. Our entire model thus is a 35-layer deep neural network with 21,348,360 trainable parameters.

Model training

We initialize the ResNet34 layers with weights trained on ImageNet[35]. To adapt the weights of the input layer from three-channel natural to two-channel UWF images, we take only the weights for the first two channels and weigh them by 3/2 to preserve the input layer's activation mass. This is the default behaviour of the PyTorch Image Models library (timm) and described in more detail at https://fastai.github.io/timmdocs/models. The linear layers of the prediction head are initialized using uniform Kaiming initialization[34] with $a = \sqrt{5}$ and PReLU αs are initialized to 0.25, both of which are the default initializations used by PyTorch.

Our neural network model is a function f parameterized by model parameters θ (weights, biases, batchnorm affine parameters and PReLU αs) that maps the image space $\mathbb{R}^{H\times W\times C}$ with height H, width W and pseudocolour-channels $C = 2$ to the eight-dimensional label space $f_\theta : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^8$. For an individual image X_i, we can then interpret the l-th element of the model output $f_\theta(X_i)_l = \hat{p}_{i,l} \in (0, 1)$ as a confidence score that X_i shows the disease indicated by the l-th label. We fit the parameters θ to minimize the total label-wise logistic loss (also referred to as binary cross-entropy) across for all training images $i \in n_{\text{images}}^{\text{train}}$. Thus, our objective is

$$\min_\theta \sum_{i=1}^{n_{\text{images}}^{\text{train}}} \sum_{l=1}^{n_{\text{labels}}} -(y_{i,l}\log(f_\theta(X_i)_l) + (1 - y_{i,l})\log(1 - f_\theta(X_i)_l))$$

where $y_{i,l} \in \{0, 1\}$ is the true value for the lth binary label of the ith image.

We optimize this objective using mini-batch stochastic gradient descent with momentum $\gamma = 0.9$. The learning rate η is set dynamically using a cosine annealing schedule with warm restarts[36], which starts with a high learning rate that is then decayed over time and reset once the minimum of the cosine function has been reached. The learning rate of a given epoch η_t is given by $\eta_t = \frac{1}{2}\left(1 + \cos\left(\frac{T_{\text{curr}}}{T_i}\pi\right)\right)\eta_{\max}$ where T_i is the number of epochs between warm restarts and T_{curr} is the number of epochs since the last restart. Once $T_{\text{curr}} = T_i$, T_{curr} is reset to 0, which implies $\eta_t = \eta_{\max}$. We use maximum learning rate $\eta_{\max} = 0.1$ and $T_i = 10$. The learning rate is updated after each mini-batch; thus, T_{curr} can take on fractional values. We train for 30 epochs with mini-batches of size 32, which was the largest size that fit our available GPU memory, shuffling the data before each epoch. Initially, we trained for 100 epochs using a constant learning rate, but after switching to a cosine schedule our model consistently converged within 30 epochs.

To make our model converge more smoothly and to better performance, we use a technique called model exponential moving average (EMA). Averaging over the trajectory in stochastic optimization has its roots in Polyak–Ruppert averaging. Other forms of averaging have been explicitly proposed in the context of modern DL[37]. However, EMA specifically appears to be a trick that is used in the literature and supported by major libraries (for example, tensorflow (https://www.tensorflow.org/api_docs/python/tf/train/ExponentialMovingAverage) and timm (https://fastai.github.io/timmdocs/training_modelEMA)), yet has not been introduced formally by a canonical work. At the start of training, we create a copy of the model parameters θ called θ_{EMA}. After the t-th mini-batch and update to θ, we update the current EMA model parameters θ_{EMA}^t to a weighted average of the previous EMA model parameters and current model parameters $\theta_{\text{EMA}}^t = \alpha\theta_{\text{EMA}}^{t-1} + (1 - \alpha)\theta$ where α is a decay hyperparameter, which we set to 0.999. At the end of training, we use θ_{EMA} as the final model parameters. When not using EMA, the label-wise validation AUCs have high variance across epochs as AUC is a ranking metric, but with EMA they converge smoothly and to higher values.

For a dataset of this size, extensive regularization is necessary to prevent overfitting. We find that even a smaller model with ResNet18 as the backbone can fit the training data perfectly (label-wise AUCs equal or almost equal to 1) after a few epochs when not applying any regularization.

A key regularization technique for computer vision is data augmentation, where images are randomly perturbed during training to yield images that retain most of their characteristic, generalizable patterns but where non-generalizable, hyper-specific patterns are more difficult to recognize. In our application, data augmentation is especially important given that our dataset contains a few thousand images. This is large by biomedical standards yet small by modern DL standards where datasets can contain hundreds of millions of images, for example, JFT-300M (https://paperswithcode.com/dataset/jft-300m).

The augmentations that we use can be split into three types: flip, domain and general augmentations. As many patterns of pathology, such as drusen, are rotation invariant, we randomly flip the images, which are all read as left eyes, horizontally with probability $p^{\text{hflip}} = 0.3$ and vertically with $p^{\text{vflip}} = 0.1$ which increases diversity while still leaving most images in the orientation we will use when predicting on unseen images. Our domain augmentations emulate some common variations in UWF image quality. We randomly scale all values in either channel with a separate float drawn uniformly from the interval [0.75, 1.25] to vary both brightness and contrast. We further apply a Gaussian blur kernel with kernel size 7×7 and σ drawn uniformly from [0.1, 1] to emulate out-of-focus images. Finally, we add Gaussian noise individually drawn from $\mathcal{N}(0, 0.1)$ to each pixel of either channel to emulate noise introduced by media opacity. All domain augmentations are applied jointly with $p = 0.9$. We also use some general augmentations that are commonly used in machine learning. We use RandomErasing[38] that with $p = 0.5$ replaces a rectangle with a fraction of the overall image area drawn uniformly from [0.05, 0.3] and aspect ratio drawn from [0.3, 3.3] with noise drawn from a standard normal distribution. We

further apply a random affine transformation, which rotates the image by an angle drawn uniformly from [−15, 15] degrees, scales it with a factor drawn uniformly from [0.8, 1.2] and shears it by an angle drawn uniformly from [−10, 10] degrees. These general augmentations are also applied jointly with $p = 0.9$ where RandomErasing is then applied with $p = 0.5$.

Further, we use mixup[29] during training. Mixup takes two images X_i and their respective target binary label vectors y_i and blends them together with linear interpolation as a new datapoint $\{X^{\text{mixup}} = \lambda X_1 + (1-\lambda)X_2, y^{\text{mixup}} = \lambda y_1 + (1-\lambda)y_2\}$. We draw the blending parameter λ from a beta distribution $\lambda \sim \beta(0.6, 0.6)$, which means that most values of λ are close to 0 or 1. The result is an image where one of the images can be seen clearly with the other one being faint, and the label vector being weighted accordingly. For instance, suppose we combine a healthy control and an image showing AMD with $\lambda = 0.9$, then the resulting image will primarily look like the healthy control with the AMD image faintly showing through, including the corresponding pathology like drusen. The target labels will be $0.9 \times 0 + 0.1 \times 1 = 0.1$ for AMD and 'any retinal disease'. Mixup reduces memorization and improves model performance in the presence of mislabelled examples and its resistance against adversarial examples[29]. During training, we pair up each input of a mini-batch with another and generate two blends, one with either input being in the first position.

Finally, we use three more regularization techniques to improve model performance. First, we apply Dropout[39] in our prediction head. Dropout regularizes the model by randomly zeroing out hidden units during training and reweighing the remaining activations to preserve mass. This prevents undesirable co-adaptation between hidden units and improves generalization. When using Dropout, we can interpret the final model during inference as an ensemble of many submodels. We use Dropout after the pooling layer with a low probability of $p = 0.1$ to encourage the model to not rely on only a few of the 512 extracted features without forcing the convolutional model to output redundant features, and after the hidden layer with $p = 0.5$, which yields the greatest diversity of submodels. Second, we use a small weight penalty of $\beta = 0.5 \times 10^{-6}$, which penalizes the L2-norm of the model parameters θ thus changing the objective to $\min_{\theta} L_{\text{original}} + \beta \parallel \theta \parallel_2$ where L_{original} is the original logistic loss to encourage lower magnitudes of each parameter, which reduces overfitting. Third, we use label smoothing (LS)[40]. LS replaces binary target labels with soft targets. This prevents the model from outputting overly confident predictions and instead encourages it to map all samples where it is highly confident to the same value, which improves model calibration[41]. We use asymmetric LS, replacing 1s with 0.99 and 0s with 0.05. This allows the model to be confident in a diagnosis while discouraging it from entirely discounting the possibility of rare disease occurring. LS can also be interpreted as encoding a belief about the possibility of images being mislabelled; we consider it more likely that a diseased image is falsely labelled as healthy control, with all labels being 0, than vice versa.

Image preprocessing

The images are shared in JPEG format at a resolution of 768 × 1,024 pixels. This is a lower resolution than the scanner acquires but still provides a detailed picture of the retina. The scanner only acquires two channels but the JPEGs also contain a third channel with low, predominantly zero values, which we assessed to have no important information. JPEG compression adds a third channel and introduces such cross-channel artefacts, and so this channel was discarded from consideration. In practice, it would be ideal to input the scans into the DL model in a lossless format, in which case no third channel would be present. We thus remove this channel for the present work and input a two-channel image into the model.

We downscale images to 384 × 512, reducing the number of pixels by a factor of 4. As the input resolution affects the size of the convolutional feature maps, this drastically lowers the GPU memory usage. We flip all right eyes horizontally, so that all images are approximately aligned. During model development, we initially experimented using a lower resolution of 256 × 341 and observed only a modest increase in performance on the validation set (overall increase in AUC < 0.005 averaged across all labels) when switching to a higher resolution of 384 × 512 (2.25 times the pixels). Previous work on a dataset of CFP images found higher image resolutions of up to 779 × 779 to be beneficial, although this benefit levelled off beyond 450 × 450 and with 1.6 million training images the dataset used was substantially larger than the TOP dataset[42]. After we selected our proposed model using the validation set and evaluated it on the test set, we retrospectively analysed whether our model would have performed better at a higher resolution. We found that using the full resolution of 768 × 1,024 increased test set AUCs by 0.0125 averaged across all labels. However, our proposed model requires less than 6 GB of GPU memory, whereas the full resolution model required more than 21.5 GB.

Benchmark models

For the non-image baseline using age and sex (the only available patient covariates), we considered three different classification algorithms: logistic regression, random forest classifier[43] and k-nearest neighbours. Logistic regression and k-nearest neighbours are standard methods and described in textbooks such as ref. [44]. As we tuned the hyperparameters of our main model, we also tune the hyperparameters of this baseline for a fair comparison. See Supplementary Section 7 for details.

For the ensemble of experts, we used the same architecture and training schedule as for our main model. However, as subsampling for rarer diseases leads to small datasets, we evaluated both the EMA and non-EMA model parameters on the validation set and picked the better-performing parameters for the test set evaluation. To obtain predictions for the general 'diseased' label from the expert binary models for each disease, we take a summary statistic of the disease-wise predictions of all expert models. We tried the minimum, mean, median and maximum and found that the maximum performed best on the validation set. For reference, using the maximum of the disease-wise predictions instead of the prediction for the 'diseased' label for our proposed DL model leads to a minor drop in performance from AUC = 0.9206 to AUC = 0.9103.

Evaluation

We evaluate our models by bootstrapping the test set 1,000 times and then reporting the mean and standard errors for each metric across the 1,000 bootstrap samples. For the internal test set where for some patients we have multiple images of the same eye, we calculate the metrics at the eye level. We think that this is the most relevant for a clinical diagnosis as some of the conditions (for example, retinal detachment) might occur in one eye but not the other and even conditions where occurrence in one eye is highly correlated with occurrence in the other eye (for example, diabetic retinopathy) can often be diagnosed by examining an individual eye. To obtain metrics at the eye level, we weigh each image during the metric calculation such that the total contribution of each eye sums to 1. For example, if a given eye was imaged three times, we assign a sample weight of 1/3 to each of them. When bootstrapping the data, we take care to calculate the correct eye weights for each bootstrap sample.

GradCAM

To generate image-level attention heat maps, we use GradCAM[20]. GradCAM creates an attention map for a given model, input image and target concept such as a class or in our case a label. We first compute the final convolutional feature map (for example, the activations of the final layer before the pooling layer) and take the gradients of the output neuron for the target concept with respect to this feature map. These gradients are then averaged across the spatial dimensions to obtain

a weight for each channel dimension of the final feature map. This is then used to take a weighted average of each spatial position of the final feature map to obtain a heat map with a single channel. Finally, negative values in this heat map are zeroed out as GradCAM aims to highlight only those parts of the image that positively contribute to predicting the target concept. The resulting GradCAM has the same resolution as the final feature map, which in the case of ResNet34 is 1/32 times the input resolution. When showing image-wise GradCAMs, we intentionally do not interpolate the heat maps to allow the reader to assess which areas are highlighted more easily. When presenting the averaged global attention maps however, the focus is not on examining the model for a specific image and there we follow established practice in the literature of using bicubic interpolation.

Progressive erasure plus progressive restoration

For PEPPR[21], we replaced erased pixels with noise drawn from a standard normal distribution so that this mirrors the RandomErasing data augmentation. Thus, our model encountered similarly erased regions during training. Furthermore, we note that we did not retrain our DL model at each step. The reported results are obtained with the weights of our final proposed DL model. This is because we want to audit what this specific model has learned and to validate that it did not make use of any shortcut artefacts[23]. Even though the model was trained on full images, it is still able to achieve very strong performance with only the central 10% of the image showing the posterior pole. To avoid data leakage, we use the global attention map obtained on the validation set, so that no test set information is used to select the regions for erasure which could, in principle, make the global attention map appear more accurate than it is. However, in practice, the global attention map for the validation and test sets look virtually identical.

Implementation

The code for this project was implemented in Python 3.8.8 and is available at https://github.com/justinengelmann/UWF_multiple_disease_detection. We used PyTorch[45] version 1.8.1 and the PyTorch Image Models (timm)[46] library version 0.4.9 for our DL models. In particular, timm was used for pretrained model weights and for model EMA and mixup. For mixup, we make a minor modification to support the multi-label case rather than the multi-class case, which we include in the code files. For general scientific computing we used NumPy[47] version 1.20.1 and for non-DL classification algorithms (for example, random forest classifier), metrics and other utility code we used scikit-learn[48] version 0.24.2. Plots were generated with the Matplotlib[49] version 3.4.2 and seaborn[50] version 0.11.1 libraries.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data are available from Hitoshi Tabuchi and the other authors of the Tsukazaki Optos Public Project subject to current export restrictions, which are imposed by Japanese legislation at the time of writing. Previously, it was publicly accessible via a project website where we obtained the copy used in this study. A subset containing images images of healthy eyes and eyes with RP used in a previous study[15] is publicly accessible directly online at https://figshare.com/authors/Masahiro_Kameoka/6020591. The external validation set we assembled from the American Society of Retina Specialists Retina Image Bank (https://imagebank.asrs.org), RetinaRocks Image Library (https://www.retinarocks.org/) and Optos Recognising Pathology resource (https://recognizingpathology.optos.com/) is described in Supplementary Section 3 in sufficient detail to reproduce the dataset. We also note that the dataset is well known within the community (for example, refs. [16,26]).

Code availability

The code for this project, a requirements.txt file listing all libraries used and their versions, and the trained model are available online at https://github.com/justinengelmann/UWF_multiple_disease_detection.

References

1. Brown, G. C. Vision and quality-of-life. *Trans. Am. Ophthalmol.Soc.* **97**, 473–511 (1999).
2. Pezzullo, L., Streatfeild, J., Simkiss, P. & Shickle, D. The economic impact of sight loss and blindness in the UK adult population. *BMC Health Serv. Res.* **18**, 63 (2018).
3. Patel, S. N., Shi, A., Wibbelsman, T. D. & Klufas, M. A. Ultra-widefield retinal imaging: an update on recent advances. *Ther. Adv. Ophthalmol.* **12**, https://journals.sagepub.com/doi/10.1177/2515841419899495 (2020).
4. Nagiel, A., Lalane, R. A., Sadda, S. R. & Schwartz, S. D. Ultra-widefield fundus imaging: a review of clinical applications and future trends. *Retina* **36**, 660–678 (2016).
5. Tan, T.-E., Ting, D. S. W., Wong, T. Y. & Sim, D. A. Deep learning for identification of peripheral retinal degeneration using ultra-wide-field fundus images: is it sufficient for clinical translation? *Ann. Transl. Med.* **8**, 611 (2020).
6. Matsuba, S. et al. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. *Int. Ophthalmol.* **39**, 1269–1275 (2019).
7. Nagasato, D. et al. Deep-learning classifier with ultrawide-field fundus ophthalmoscopy for detecting branch retinal vein occlusion. *Int. J. Ophthalmol.* **12**, 94–99 (2019).
8. Nagasato, D. et al. Deep neural network-based method for detecting central retinal vein occlusion using ultrawide-field fundus ophthalmoscopy. *J. Ophthalmol.* **2018**, 1875431 (2018).
9. Tabuchi, H., Masumoto, H., Nakakura, S., Noguchi, A. & Tanabe, H. Discrimination ability of glaucoma via DCNNs models from ultra-wide angle fundus images comparing either full or confined to the optic disc. In *Asian Conference on Computer Vision* 229–234 (Springer, 2018).
10. Masumoto, H. et al. Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. *J. Glaucoma* **27**, 647–652 (2018).
11. Nagasawa, T. et al. Accuracy of deep learning, a machine learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting idiopathic macular holes. *PeerJ* **6**, e5696 (2018).
12. Nagasawa, T. et al. Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatment-naive proliferative diabetic retinopathy. *Int. Ophthalmol.* **39**, 2153–2159 (2019).
13. Ohsugi, H., Tabuchi, H., Enno, H. & Ishitobi, N. Accuracy of deep learning, a machine-learning technology, using ultra–wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment. *Sci. Rep.* **7**, 9425 (2017).
14. Masumoto, H. et al. Retinal detachment screening with ensembles of neural network models. In *Asian Conference on Computer Vision* 251–260 (Springer, 2018).
15. Masumoto, H. et al. Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. *PeerJ* **7**, e6900 (2019).
16. Antaki, F. et al. Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. *Br. J. Ophthalmol.* https://bjo.bmj.com/content/early/2021/08/02/bjophthalmol-2021-319030.info (2021).
17. Hemelings, R. et al. Deep learning on fundus images detects glaucoma beyond the optic disc. *Sci. Rep.* **11**, 20313 (2021).

18. Duker, J. S. et al. The international vitreomacular traction study group classification of vitreomacular adhesion, traction, and macular hole. *Ophthalmology* **120**, 2611–2619 (2013).
19. Beede, E. et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2020).
20. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision* 618–626 (IEEE, 2017).
21. Engelmann, J., Storkey, A. & Bernabeu, M. O. Global explainability in aligned image modalities. Preprint at https://arxiv.org/abs/2112.09591 (2021).
22. Wilkinson, C. P., Hinton, D. R., Sadda, S. R. & Wiedemann, P. *Ryan's Retina* 6th edn (Elsevier Health Sciences, 2018).
23. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
24. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
25. Yamashita, T. et al. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transl. Vis. Sci. Technol.* **9**, 4 (2020).
26. Khan, S. M. et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit. Health* **3**, e51–e66 (2021).
27. González-Gonzalo, C., Liefers, B., van Ginneken, B. & Sánchez, C. I. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images. *IEEE Tran. Med. Imaging* **39**, 3499–3511 (2020).
28. Quellec, G., Charrière, K., Boudi, Y., Cochener, B. & Lamard, M. Deep image mining for diabetic retinopathy screening. *Med. Image Anal.* **39**, 178–193 (2017).
29. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization. Preprint at https://arxiv.org/abs/1710.09412 (2017).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
31. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* 448–456 (PMLR, 2015).
32. Wightman, R., Touvron, H. & Jégou, H. ResNet strikes back: an improved training procedure in timm. Preprint at https://arxiv.org/abs/2110.00476 (2021).
33. Bello, I. et al. Revisiting ResNets: improved training and scaling strategies. Preprint at https://arxiv.org/abs/2103.07579 (2021).
34. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proc. IEEE International Conference on Computer Vision* 1026–1034 (IEEE, 2015).
35. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
36. Loshchilov, I. & Hutter, F. SGDR: stochastic gradient descent with warm restarts. Preprint at https://arxiv.org/abs/1608.03983 (2016).
37. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. Averaging weights leads to wider optima and better generalization. Preprint at https://arxiv.org/abs/1803.05407 (2018).
38. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 34, 13001–13008 (Association for the Advancement of Artificial Intelligence, 2020).
39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
41. Müller, R., Kornblith, S. & Hinton, G. When does label smoothing help? Preprint at https://arxiv.org/abs/1906.02629 (2019).
42. Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
43. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
44. Friedman, J. et al. *The Elements of Statistical Learning* Vol. 1 (Springer Series in Statistics, Springer, 2001).
45. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
46. Wightman, R. PyTorch image models. *GitHub* https://github.com/rwightman/pytorch-image-models (2019).
47. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
48. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sc. Eng.* **9**, 90–95 (2007).
50. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Soft.* **6**, 3021 (2021).

Acknowledgements

We thank H. Masumoto and H. Tabuchi as well as D. Nagasato, S. Nakakura, M. Kameoka, R. Aoki, T. Sogawa, S. Matsuba, H. Tanabe, T. Nagasawa, Y. Yoshizumi, T. Sonobe, T. Yamauchi and all their colleagues at Tsukazaki Hospital for releasing the TOP dataset. This is a great contribution to AI research in ophthalmology for which we are most grateful. We also thank the American Society of Retina Specialists for their Retina Image Bank, and RetinaRocks for their Image Library. We further thank all users that submitted images for research use to these online repositories or elsewhere. This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. This work was supported by The Royal College of Surgeons of Edinburgh, Sight Scotland, The RS Macdonald Charitable Trust, Chief Scientist Office, and Edinburgh & Lothians Health Foundation through a proof-of-concept award for the SCONe project. Grant EP/S02431X/1: J.E. SCONe project grants: A.D.M. and E.P.

Author contributions

J.E. was responsible for all aspects of this work, including conceptualization, study design/methods, experiments, analysis, interpretation, figures and writing. A.S. and M.O.B. jointly supervised and contributed to all aspects of this work. A.D.M., I.J.C.M. and E.P. provided domain expertise regarding ophthalmology and ultra-widefield imaging, assessed the top 20 false positives, and provided feedback on the interpretation of the results.

Competing interests

The authors declare no competing interests.

Additional information

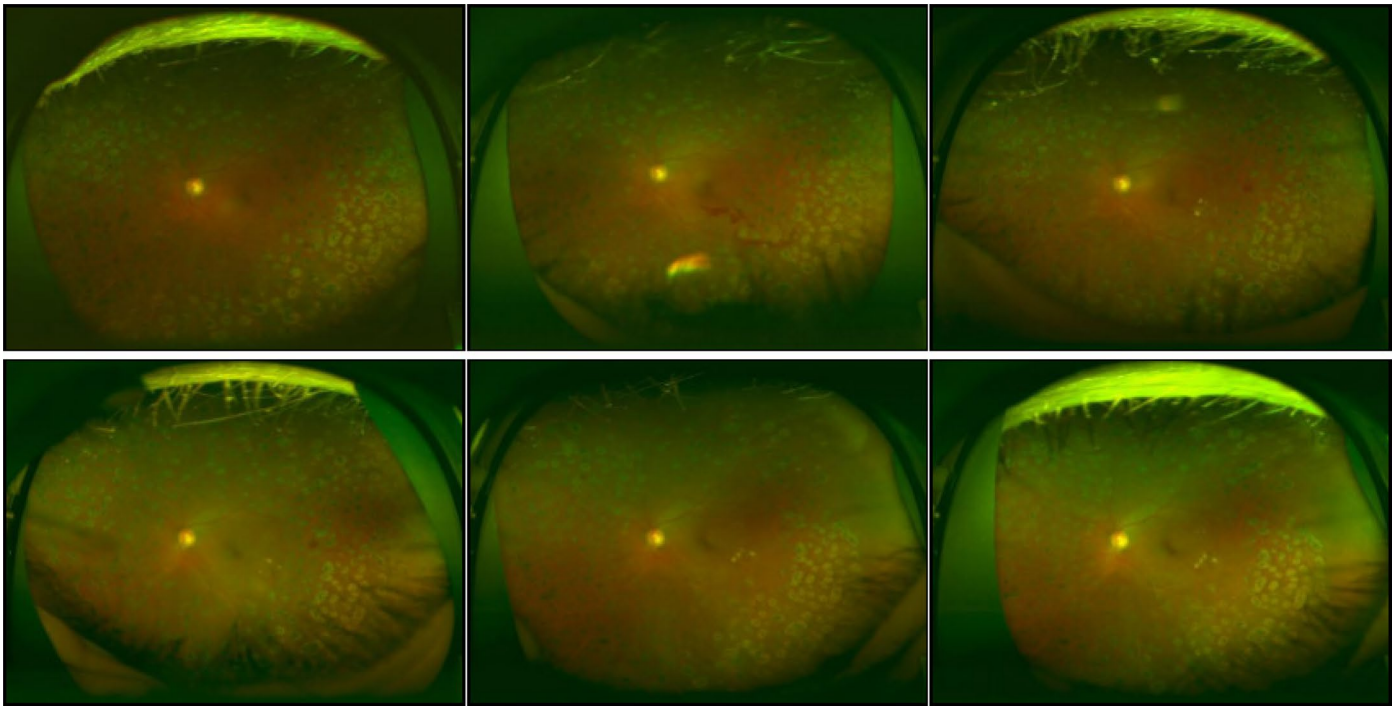
Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00566-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00566-5.

Correspondence and requests for materials should be addressed to Justin Engelmann or Miguel O. Bernabeu.

Peer review information *Nature Machine Intelligence* thanks Edward Korot and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-022-00566-5



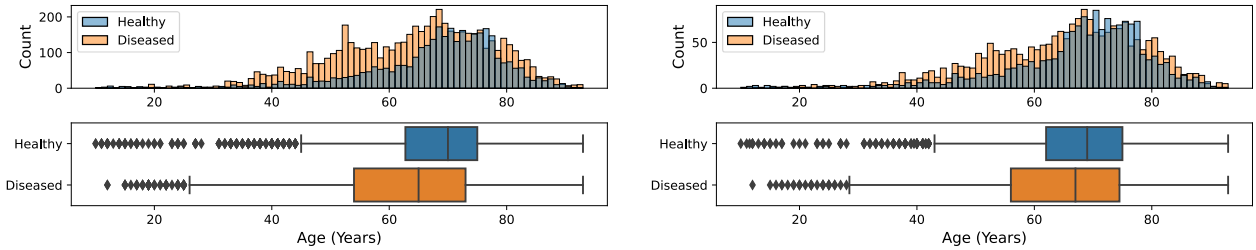
Extended Data Fig. 1 | Six of nine images showing the same eye of the same patient. Six of nine images showing the same eye of the same patient. All images show DR according to the labels. While there are some differences between the images in terms of artefacts and pathology, the general pattern of the pathology is consistent between images and could be memorized by a model.

S Supplementary Materials

S.1 Exploratory data analysis of the TOP dataset

The TOP dataset contains a label for diabetes mellitus (DM) as determined by a blood test with no further details given. We find that this label and DR co-occur often (Dice coefficient of 0.09). However, in general not all patients with DM will have DR, so the high co-occurrence in the TOP dataset might indicate that a majority of DM patients were examined because of suspected DR or that DM blood tests were done primarily for patients showing signs of DR. Curiously, 39 images showed DR but the patient did not have DM according to the labels, which supports the idea that the DM label refers only to blood tests done at Tsukazaki hospital and thus there might be patients with DM who are recorded as DM negative in the dataset.

The equal sex balance which is surprising given that patients in the TOP dataset tend to be of advanced age and that females tend to live longer. One possible explanation could be a conscious decision by the researchers at Tsukazaki hospital to include an equal proportion of males and females. Thus, it could be a sign of possible selection bias.



(a) Image-level (n=9121) **(b)** Patient-level (n=3763)

Figure S1: Distribution of patient age stratified by disease status for the train set. Bin width is set to 1, the granularity of age in the dataset. To define the disease status at the patient-level, we classify a patient as diseased if any of their images showed a retinal disease or if they have DM.

More detailed analysis was done on the train set only after the data was split. This is to ensure that we do not leak information from the test set into our modelling process. Fig. S1 shows the distribution of age stratified by disease status. There is some coverage of almost all age groups, from young children to centenaries. Most patients are between 50 and 80 which is likely reflective of the fact that most retinal diseases occur in older patients and of Japan's demographic makeup. Surprisingly, we find that healthy patients are generally older than diseased patients. This difference is more apparent at the image-level but persists at the patient-level, even when using the sweeping definition of classifying a patient as diseased if they had any disease associated with any of their images. We think that this is another sign of possible selection bias. Patients get examined at a specialist clinic for a reason. With younger patients, this would be concrete reasons to suspect retinal diseases like self-reported deteriorating vision, whereas older patients might get examined routinely due to their age, even if they have no symptoms.

Another possible sign of selection bias is that the labels for DM and DR have low co-occurrence with the other retinal diseases in the TOP dataset. We would generally expect that patients with diabetes have an increased risk of many other conditions, including retinal diseases other than DR. However, looking at the TOP dataset, patients with DR/DM appear to have a much lower risk of having other retinal diseases than patients without DR/DM. This could again be due to patients being referred to Tsukazaki hospital for a reason, which could be suspected DR or some other reason. Patients that do not have DR/DM are more likely to have been referred for some other reason and thus have other retinal diseases more often.

S.2 Overview of work by researchers from Tsukazaki hospital

Table S1: An overview of prior work on the TOP dataset conducted by researchers from Tsukazaki hospital. "Custom 1" is an architecture consisting of 3 convolutional layers, followed by a maxpool and a linear prediction head. "Custom 2" consists of 3 convolutional layers each followed by a maxpool, followed by a prediction head containing one hidden layer. "Ensemble" refers to an ensemble of multiple convolutional neural networks. Where multiple tasks where considered (e.g. distinguishing a subtype from healthy controls), we selected the most relevant performance metric. One work did not report AUC, so sensitivity and specificity are reported instead.

Reference	Disease	Diseased	Controls	Reported Performance	Model	Validation	GradCAM?
[6]	AMD	137	227	AUC=0.9976	Custom 1	70-30 split	Yes
[7]	RVO	237	229	AUC=0.976	VGG16	k-fold CV	Yes
[8]	RVO	125	238	AUC=0.989	VGG16	k-fold CV	Yes
[9]	Gla	950	1677	AUC=0.987	VGG16	k-fold CV	No
[10]	Gla	982	417	AUC=0.872	Custom 1	80-20 split	Yes
[11]	MH	195	715	AUC=0.9993	Custom 1	80-20 split	Yes
[12]	DR	132	246	AUC=0.969	VGG16	k-fold CV	Yes
[13]	RD	411	420	AUC=0.988	Custom 2	75-25 split	No
[14]	RD	600	818	Sensitivity=0.973, Specificity=0.915	Ensemble	k-fold CV	No
[15]	RP	150	223	AUC=0.998	VGG16	k-fold CV	Yes

S.3 Details of the external test set

Table S2 lists the sources, filenames, labels and predicted probabilities for all images of the external test set. Using these data sources to assemble an external test set was inspired by recent work by [16]. "ORP" refers to Optos'© Recognizing Pathology which contains UWF images with clinical labels designed as "a searchable reference resource to support clinical decision making". "ASRS" refers to the American Society of Retina Specialists' Retina Image Bank® and "RetinaRocks" for the RetinaRocks Image Library.

For images where no clear label is provided at the source we take the labels from [16] who are trained ophthalmologists. We thank everyone who made these images available for research use, for which we are most grateful. We do not reproduce any images here, but we note that the American Society of Retina Specialists requests the following acknowledgement for each image: "This image was originally published in the Retina Image Bank® website. © the American Society of Retina Specialists." The exact submitter for each image from that database can be found at the respective URL. We do not print the full links here due to space constraints, but they are available as hyperlinks in the "Source/URL" column. RetinaRocks images are provided in a Google Drive and thus cannot be linked directly. The filename allows to identify the exact image in this case.

We were unable to find external UWF images that were described as showing a Macular Hole. For Glaucoma, we were only able to find one additional image. In the future, we hope to be able to test our model on larger external datasets.

For stereo images from ORP, we follow [16] in taking the left of the two panels to avoid cherry-picking. Some images from RetinaRocks had both eyes side by side in a single image. For those, we simply manually split them into two separate files. We applied our standard data pipeline to all external images: We resized them to the same resolution we used for the TOP dataset, removed the third channel present in JPG images, and flipped right eyes horizontally. No further processing was done, e.g. to correct for the different scale of cropped images, different aspect ratios, or watermarks. Together with the fact that many of these images were taken with different UWF scanners than what

was used at Tsukazaki hospital, this makes the external dataset a very challenging stress test for our model. Furthermore, we want to note that any particularities in the data collection that might be present in the TOP dataset will also be absent from these images taken from a wide variety of sources. Thus, good performance on this dataset would indicate good generalisation of a model.

Table S2: Details of the external test set.

Image	Label	Source/URL	Filename	$\hat{p}(diseased)$
1	AMD	ORP	Color-Atrophic-AMD-with-Geographic-Atrophy-OD-California.jpg	0.982783
2	AMD	ORP	Color-Dry-AMD-OD-California-1.jpg	0.937082
3	AMD	ORP	Color-Dry-AMD-OS-California-1.jpg	0.941377
4	AMD	ORP	California-AMD.jpg	0.873758
5	AMD	ORP	California-AMD-Dry.jpg	0.923850
6	AMD	ORP	California-AMD-3.jpg	0.951265
7	AMD	ORP	California-AMD-2.jpg	0.982928
8	AMD	ORP	Daytona-Projected-AMD-2-of-2-1.jpg	0.881232
9	AMD	ORP	Daytona-Projected-AMD-1-of-2.jpg	0.952176
10	AMD	ORP	Daytona-Projected-Wet-AMD-1.jpg	0.844429
11	AMD	ORP	Color-Wet-AMD-OD-California.jpg	0.936324
12	AMD	ORP	Color-Wet-AMD-OS-California.jpg	0.758085
13	AMD	ORP	P200Tx-Projected-AMD-Geographic-Atrophy.jpg	0.830430
14	AMD	ORP	AMD-Wet-California-Courtesy-Tim-Steffens-CRA-OCT-C-FOPS.jpg	0.987892
15	AMD	ORP	Color-Atrophic-AMD-with-Geographic-Atrophy-OS-California.jpg	0.936971
16	DR	ORP	P200Tx-Projected-Severe-NPDR-with-DME-HariprasadDiabeticRetinopathy22.jpg	0.869121
17	DR	ORP	Color-NPDR-with-Macular-Edema-OD-California.jpg	0.907351
18	DR	ORP	P200Tx-Projected-Diabetic-RetinopathyDiabeticRetinopathy16.jpg	0.886153
19	DR	ORP	Color-NPDR-with-Macular-Edema-OS-California.jpg	0.970982
20	DR	ORP	Daytona-Projected-PDR-with-PDP-1DiabeticRetinopathy11.jpg	0.847240
21	DR	ORP	P200Tx-Projected-DR-with-PRP-1DiabeticRetinopathy17.jpg	0.940622
22	DR	ORP	California-DR-and-PRP-StangaDiabeticRetinopathy4.jpg	0.928405
23	DR	ASRS	ASRS-RIB-Image-27747.jpg	0.955919
24	DR	ORP	Color-Severe-NPDR2-California.jpg	0.649301
25	DR	ORP	California-Proliferative-Diabetic-RetinopathyDiabeticRetinopathy8.jpg	0.963178
26	DR	ORP	Color-PDR2-OS-California.jpg	0.942234
27	DR	ORP	Color-Severe-NPDR-California.jpg	0.970263
28	DR	ORP	Color-PDR2-OD-California.jpg	0.933320
29	DR	ORP	California-DR-Stanga-1DiabeticRetinopathy5.jpg	0.948500
30	DR	ORP	California-DR-and-PRP-Sadda-1DiabeticRetinopathy1.jpg	0.967951
31	DR	ORP	P200Tx-Projected-Severe-NPDR-with-DME-Hariprasad-2DiabeticRetinopathy21.jpg	0.954090
32	DR	ORP	P200Tx-Projected-Diabetic-Retinopathy-with-Mac-Grid-1DiabeticRetinopathy14.jpg	0.949963
33	DR,AMD	ORP	Color-Wet-AMD-and-NPDR-California-Courtesy-Mandar-Joshi-MD_result-1.jpg	0.968636
34	DR,AMD	ORP	Color-Wet-AMD-and-NPDR-OS-California-Courtesy-Mandar-Joshi-MD_result-1.jpg	0.945011
35	DR,Drusen(AMD?)	ORP	Daytona-Projected-DR-Peripheral-DrusenDiabeticRetinopathy10.jpg	0.961227
36	Gla	ORP	Glaucoma-Suspect-California-William-Lesko-MD-North-Jersey-Eye-Associat..._result.jpg	0.823686
37	Healthy	ORP	Color-Healthy-California-Courtesy-Michael-Singer-MD_result.jpg	0.844392
38	Healthy	ORP	Color-Stereo-Healthy-OS-California_leftpanel.jpg	0.727054
39	Healthy	ORP	California-optomap-am-Healthy.jpg	0.747208
40	Healthy	Web	Technology-1.jpg	0.610550
41	Healthy	Web	Color-Fundus1.jpg	0.445265
42	Healthy	ORP	Color-Stereo-Healthy-OD-California_leftpanel.jpg	0.769579
43	Healthy	RetinaRocks	Normal HOZ-20190131 (1).jpg	0.578932
44	Healthy	ORP	Daytona-Projected-Healthy-Retina-Adult.jpg	0.667379
45	Healthy	ORP	Color-Healthy-P200Tx.jpg	0.494873
46	Healthy	ORP	Color-Healthy-Child-P200Tx.jpg	0.716572
47	Healthy	ORP	Color-Healthy-California.jpg	0.581765
48	Healthy	ASRS	ASRS-RIB-Image-78895.jpg	0.684483
49	RD	ASRS	ASRS-RIB-Image-30011.jpg	0.920010
50	RD	RetinaRocks	RRD NVG-20191210.jpg	0.939834
51	RD	RetinaRocks	RRD GSV-20190213 (1).jpg	0.922635
52	RD	RetinaRocks	RRD WVI-20191105 (1) Recurrent with PVR and new small temporal hole.jpg	0.945073
53	RD	ASRS	ASRS-RIB-Image-26484.jpg	0.935745
54	RD	ASRS	ASRS-RIB-Image-25718.jpg	0.940452
55	RD	RetinaRocks	RRD SLO-20131219 Giant tear.jpg	0.901568
56	RP	RetinaRocks	RP LYV2-20181121 (5) _ LEFTEYE.jpg	0.974408
57	RP	RetinaRocks	RP LYV2-20181121 (5) _ RIGHTEYE.jpg	0.958171
58	RP	RetinaRocks	RP DVY-20190416 (2).jpg	0.970836
59	RP	ASRS	ASRS-RIB-Image-18045.jpg	0.964437
60	RP	Reddit	hij0f9pkqn441.jpg	0.952194
61	RP	ASRS	ASRS-RIB-Image-28324.jpg	0.928548
62	RP	RetinaRocks	RP GSL1-20190731 (1) 20-30 OU With OCT and Optos with FAF _ RIGHTEYE.jpg	0.973185
63	RP	RetinaRocks	RP DVY-20190416 (1) From SCO.jpg	0.958874
64	RP	ASRS	ASRS-RIB-Image-27209.jpg	0.912519
65	RP	RetinaRocks	RP GSL1-20190731 (1) 20-30 OU With OCT and Optos with FAF _ LEFTEYE.jpg	0.970093
66	RVO	ORP	P200Tx-Projected-Retinal-Vein-Occlusion-with-Ozurdex-injectionOcclusion31.jpg	0.816859
67	RVO	RetinaRocks	BRVO Major URI-20200106 Extramacular.jpg	0.932180
68	RVO	RetinaRocks	Unknown DSV-20190618 (1) 74YOF Possible multifocal BRVO's.jpg	0.930758
69	RVO	RetinaRocks	CRVO Ischemic PVM-20190709 (1) HM OD 20-50 OS Fresh CRVO OD Old resolved CRVO OS.jpg	0.925773
70	RVO	ORP	P200Tx-Projected-Retinal-Vein-OcclusionOcclusion33.jpg	0.947042
71	RVO	ASRS	ASRS-RIB-Image-66298.jpg	0.946579
72	RVO	ORP	Color-BRVO-California.jpg	0.960900
73	RVO	ORP	Color-Hemi-Retinal-Vein-Occlusion-California.jpg	0.944914
74	RVO	RetinaRocks	BRVO Major SZB-20191203.jpg	0.929029
75	RVO	ASRS	ASRS-RIB-Image-25606.jpg	0.889695

S.4 Model generalises to unseen disease (held-out images showing Artery Occlusion)

We evaluated our model on the 21 images showing Atery Occlusion (AO). We had excluded them as including a label for a disease with only 21 available images would have been unlikely to be useful, particular when using a three-way data split on the patient-level. However, this allows us now to evaluate the model on an "unseen" disease. 11 of those 21 images show only AO, 10 also show another disease. As our model might just recognise the other diseases, we stratify our analysis accordingly.

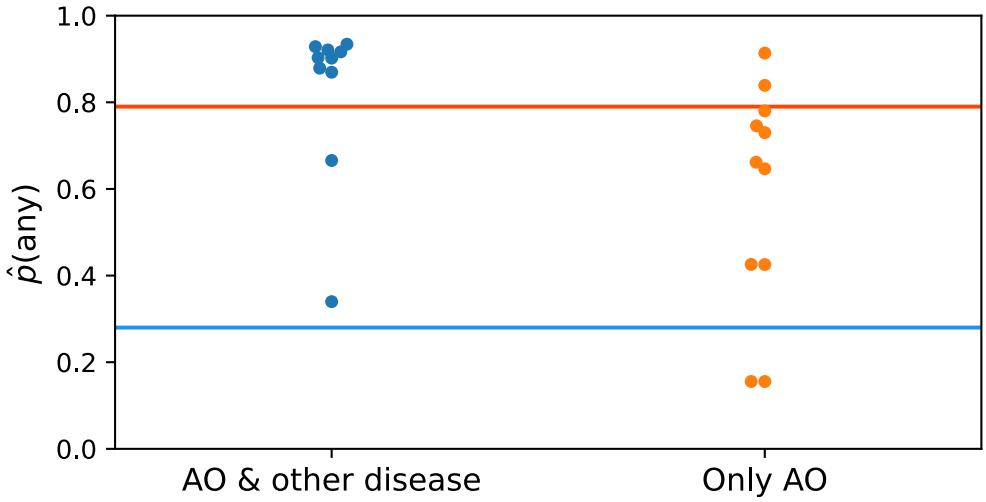
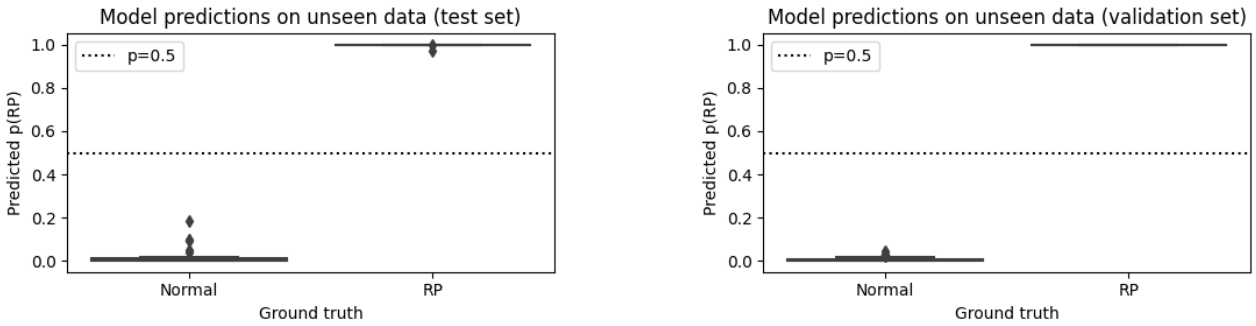


Figure S2: Results of evaluating our model on the 21 held-out AO images showing the predicted probability of being diseased $\hat{p}$(diseased) stratified by whether the images had another condition apart from AO. The red and blue horizontal lines plot indicate the conservative threshold $\hat{p}^t_{\text{conservative}} = 0.79$ and and less conserviative threshold $\hat{p}^t_{\text{less conservative}} = 0.28$, respectively.

We find that of 11 images that only show Artery Occlusion, our model correctly identifies 9 as diseased at the less conservative threshold but only 2 at the conservative threshold. The remaining 10 images show Artery Occlusion and a further disease, and our model identifies all 10 as diseased at the less conservative threshold, and 8 at the conservative incidence threshold. This is encouraging performance and highlights the value of the label for "diseased" as for some images, our model was not confident in any of the seven diseases that it was trained on, but nontheless very confident that there is some disease present. The generalisability to unseen diseases would be a very valuable property for practical applications, provided that the model does not frequently falsely flag up healthy patients.

S.5 Performance on subset of data used in a previous study on RP

We briefly experimented with a subset of 223 healthy controls and 150 images showing RP that were used in prior work on the TOP dataset [15]. We split the into train, validation and test sets containing 70, 15 and 15% of the images. We find that even when using a ResNet18 with a linear prediction head, and only simple flip augmentations as regularisation, we can easily achieve perfect separation between the two classes on both the validation and test set with very high confidence in the correct labels (implying very low Brier scores). This matches what has been reported in the literature [16]. This perfect separation (AUC=1), achieved easily, with little tuning or effort in designing the model, contrasts with the very high yet not perfect separation our model obtained in our work (AUC=0.9438).



(a) Model performance on the test set.

(b) Model performance on the validation set.

Figure S3: Results of evaluating a simpler, binary model on a clean subset of RP images that was used in prior work [15].

We see this as an indication that our methodological improvements lead to a more realistic estimate of model performance.

S.6 Assessment of Top 20 False Positives

Table S3: Results of the assessment of the top 20 false positives. The scores indicates the grader's assessment whether the image seems to show pathology according to the following scale: -2: Definitely shows no pathology -1: Probably shows no pathology 0: Unclear 1: Probably shows pathology 2: Definitely shows pathology

False positive	Filename	Score AM	Score IM	Score EP	Median score
1	003627_01.jpg	2	2	2	2
2	003626_05.jpg	2	2	2	2
3	002946_02.jpg	2	2	2	2
4	002947_00.jpg	1	2	2	2
5	000053_01.jpg	2	2	2	2
6	000440_00.jpg	0	1	2	1
7	003627_00.jpg	2	2	2	2
8	005008_02.jpg	2	2	1	2
9	000054_01.jpg	1	0	1	1
10	003626_04.jpg	2	2	2	2
11	001984_00.jpg	-1	-1	0	-1
12	000686_01.jpg	0	-1	0	0
13	001389_01.jpg	0	-1	1	0
14	002313_00.jpg	2	2	2	2
15	001778_00.jpg	-2	-1	1	-1
16	001436_00.jpg	2	2	2	2
17	001389_00.jpg	-1	-1	-1	-1
18	003376_01.jpg	1	1	1	1
19	004733_00.jpg	1	2	1	1
20	001566_00.jpg	2	-1	0	0
Median		1.5	2	1.5	1.5
Mean		1	0.9	1.25	1.05
Count (score>0)		14	13	16	14

We assessed the top 20 most confident false positives our model generated on the test set, i.e. we

selected the 20 images with the highest $\hat{p}(diseased)$ where the image had no disease according to the labels. This allows us to get a sense of the label noise in the TOP dataset.

The three co-authors with relevant experience each assessed these images independently. A.M. and I.M. are experienced clinicians and researchers in ophthalmology. E.P. is a researcher in retinal image analysis and completed her PhD on automated analysis of UWF images.

The instructions for the assessment were as follows: *"Please briefly evaluate whether each of the images seems to show pathology and assign a score from -2 to 2 according to the following scale. -2: Definitely shows no pathology -1: Probably shows no pathology 0: Unclear 1: Probably shows pathology 2: Definitely shows pathology"*

In section S.6, we report the results of this assessment. For 14 of the images, the median score was greater than 0, indicating that a majority of the graders thought that the image might show pathology. This suggests that these images are likely to show pathology. For 17 of the images, at least one of the graders thought that the image might show pathology. This assessment is limited as we only examined a small number of images. Furthermore, they were assessed by co-authors of this work. Despite our efforts to remain impartial, this could have introduced bias into the assessment.

S.7 Age+Sex benchmark model algorithm and hyperparameter selection

For Logistic Regression, we min-max scale input features to $[0, 1]$ and choose the penalty type from $\{L1, L2, \text{ElasticNet}\}$ and inverse penalty strength C from $\{10, 1, 0.1\}$. ElasticNet is a combination of L1 (LASSO) and L2 (Ridge) penalties, which we weigh equally. For RFC, we grow 100 trees and choose number of features per tree from $\{\sqrt{n_{\text{features}}}, 0.1n_{\text{features}}, n_{\text{features}}\}$ and maximum tree depth from $\{6, 10, \text{Unlimited}\}$. For KNN, we choose the number of neighbours to consider k from $\{5, 15, 30, 60\}$ and the distance measure used from $\{\text{Manhattan}, \text{Euclidean}\}$. We considered all classifiers and hyperparameter settings and selected the combination that performed best in terms of AUC on distinguishing unhealthy from healthy images on the validation set, the same criterion used for selecting the final DL model.

S.8 Patients, eyes, and images per disease

Table S4: Detailed overview over the number of patients/eyes/images per disease in the whole dataset and our three subsets. Note that here we report how many patients/eyes/images are labelled as showing a particular disease. As even the same image can show multiple diseases, this means that this table intentionally double counts. Table 1, on the other hand, counts every patient exactly once according to the procedure we describe.

	TOP Dataset			Train Set			Validation Set			Test Set		
	Patients	Eyes	Images	Patients	Eyes	Images	Patients	Eyes	Images	Patients	Eyes	Images
Disease												
MH	185	188	222	129	131	156	27	27	30	29	30	36
RP	111	202	258	78	139	172	17	31	39	16	32	47
AMD	292	377	413	205	260	285	44	57	61	43	60	67
RVO	528	576	771	370	406	542	80	87	121	78	83	108
RD	452	459	974	321	326	713	66	67	129	65	66	132
Gla	999	1679	2619	703	1177	1831	146	249	386	150	253	402
DR	749	1288	3320	521	902	2338	113	194	494	115	192	488

S.9 Non-bootstrapped performance numbers

Table S5: Test set performance (AUC) of baselines and final model for each label. AUC assesses how well a model can separate positive and negative samples for a given label. Images are weighted such that each eye has a total weight of 1, even if a specific eye was imaged multiple times. Higher is better, best values in bold.

	Diseased	DR	Gla	RD	RVO	AMD	RP	MH
Logistic Regression with Age + Sex	0.5964	0.5988	0.5155	0.7676	0.4892	0.8021	0.6776	0.5625
Ensemble of Experts (binary DL models + balanced data)	0.8318 *	0.8432	0.9141	0.9217	0.8996	0.7113	**0.9490**	0.6454
Ours (Single multi-label DL model + realistic data)	**0.9206**	**0.9125**	**0.9422**	**0.9753**	**0.9468**	**0.9510**	0.9438	**0.7987**

* Using maximum of individual predictions (Section 4.5).

Table S6: Test set performance (Brier score) of baselines and final model for each label. Brier score is sensitive to how well a model's predicted probabilities are calibrated. Images are weighted such that each eye has a total weight of 1, even if a specific eye was imaged multiple times. Lower is better, best values in bold.

	Diseased	DR	Gla	RD	RVO	AMD	RP	MH
Logistic Regression with Age + Sex	0.2522	0.1354	0.1580	0.0466	0.0567	0.0426	0.0242	**0.0227**
Ensemble of Experts (binary DL models + balanced data)	0.1919 *	0.1421	0.1182	0.0780	0.1211	0.2086	0.0993	0.2373
Ours (Single multi-label DL model + realistic data)	**0.1144**	**0.0690**	**0.0645**	**0.0150**	**0.0283**	**0.0269**	**0.0081**	0.0238

* Using maximum of individual predictions (Section 4.5).

The explainability method used to generate the global attention heatmaps in the paper has been described in a separate report that is available as a pre-print on the arXiv repository (Engelmann et al., 2021). It has been presented at the Interpretable Machine Learning in Healthcare workshop at the 2022 International Conference on Machine Learning, where it was peer-reviewed prior to acceptance at the workshop. However, this workshop is considered "non-archival" meaning that the work could later be submitted to a proper journal or conference with proceedings and might not yet be final. Thus, I do not consider this report to be a full paper and did not include it in this thesis.

2.3 Conclusion

In this paper, we presented a deep learning model that achieved very promising performance - in internal and external validation - while framing the problem in a much more challenging and realistic way compared to previous work on the same dataset - namely dealing with multiple key diseases without excluding difficult cases. We further use methods for explainability and find that the model highlights regions of pathology relevant to its predictions, and that "globally", i.e. at a dataset level, the regions for each disease broadly aligned with where we would expect the corresponding pathology to occur. Generally, the posterior pole, the central region of the ultra-widefield images, were the most important. This is what we would expect and finding this in a purely data-driven way serves as a useful sanity check.

One finding that is more surprising is that performance does not drop substantially when removing 90% of the image that were least important for our model, according to our global attention maps. In other words, the model retained comparable performance to when it received the full images, even when we erased everything but the posterior pole. It is important to note that these evaluations were conducted with a single trained model, rather than a different model that was trained on images with everything but the posterior pole erased. This raises some questions regarding the clinical utility of ultra-widefield imaging, at least as far as automated disease detection is concerned.

However, this work has a number of weaknesses. The external validation data is relatively small and from a variety of sources, particularly images shared online by clinicians. There might be considerable selection bias here: images of especially severe cases might be shared more often, and more severe cases are easier to classify. On

the other hand, the external validation images contains data from newer Optos devices as well as a few Zeiss Clarus ultra-widefield images, both of which are characteristically different in appearance from the images used for training the model. They are also from different countries, populations, and settings. Thus, there are factors that could make the external validation dataset particularly easy or particularly challenging. Still, we can only draw very limited conclusions regarding the generalisability of our model from it.

Another limitation is that the labels for this dataset were only binary, with no information relating to severity (e.g. diabetic retinopathy grade) or subtype (e.g. wet or dry age-related macular degeneration). This limits the clinical utility of our model and also has two important implications regarding the finding that the posterior pole itself gave comparable performance to the full ultra-widefield images: First, maybe the periphery is not important for disease detection but would be important to judge severity or subtype. Second, maybe most cases in the dataset were of severe disease where pathology is present everywhere, including in the posterior pole - yet for less severe disease, e.g. non-proliferative diabetic retinopathy, the periphery would be important even for detection. To investigate the utility of ultra-widefield imaging compared to standard field colour fundus images, we would ideally have a dataset of both modalities being acquired of the same eyes during the same visit. Finally, the data is from a single hospital with some selection biases, e.g. older patients had less disease as discussed in the paper, and of an older Optos device that has since been discontinued.

In the future, these limitations should be addressed by developing a new model on a more comprehensive dataset using a recent, commercially available device which is then externally validated in larger, cohesive datasets. This is work I am currently undertaking with colleagues from Japan and England. There are also interesting questions the paper raises that I have not yet touched on. For example, the performance for detecting macular holes and glaucoma is quite high, which surprises some of my clinical collaborators who think those are difficult to assess from ultra-widefield images alone. Yet, in the currently on-going work using a newer device where we have a larger dataset and labels that were carefully adjudicated, we still get a similar level of performance for these conditions (preliminary results, data not shown here). Thus, it would be worth investigating what features the deep learning model uses and whether this might yield some new insight into the respective pathologies or their appearance on ultra-widefield images. Since the publication of the paper, I have been trying to identify a dataset of paired standard field colour fundus and ultra-widefield fundus images of

the same eyes during the same visit, but to date I have been unable to access such a dataset. Finally, I am in the early stages of other projects relating to analysing ultra-widefield images with machine learning for specific conditions to support research by clinical collaborators of mine.

Chapter 3

Machine learning for robust and efficient computation of retinal fractal dimension from colour fundus images: deep approximation of retinal traits (DART)

3.1 Introduction

Retinal traits that quantify aspects of interest from retinal images are investigated in relation to systemic health, a field of study sometimes referred to as “oculomics” (Wagner et al., 2020). The hope is that retinal images, which are widespread, non-invasive, fast-to-acquire, and comparatively low cost, could provide information about systemic health and might one day allow for better risk prediction of systemic conditions. A particularly promising retinal trait is retinal fractal dimension computed from colour fundus images which captures the complexity of the retinal vasculature and has been investigated in the context of cardiovascular (Cheung et al., 2012; Ify Mordi and Emanuele Trucco, 2022; Villaplana-Velasco et al., 2023; Zekavat et al., 2022) and neurovascular (Lemmens et al., 2020; Luben et al., 2022) disease.

Retinal fractal dimension is computed using retinal image analysis pipelines that traditionally first segment the blood vessels, then refine the segmentation e.g. by skeletonising it or removing the vessels in the optic disc, and then compute fractal dimension using methods such as box counting (Huang et al., 2016; Stosic and Stosic, 2006). Examples of such pipelines are VAMPIRE (Trucco et al., 2013) and AutoMorph (Zhou et al., 2022). Fractal dimension in general is a measure of complexity of an object, so

in the context of retinal vasculature it captures how complex the vessel structure is. Lower complexity in the vessel structure might be a sign of poorer vessel health, and poorer vessel health in the retina might be a proxy for poorer vessel health elsewhere in the body such as the heart or brain. However, pipelines for computing fractal dimension require good image quality and in research datasets like UK Biobank between a quarter (Zekavat et al., 2022) and close to half of the images (MacGillivray et al., 2015; Villaplana-Velasco et al., 2023) might be excluded from analysis. This is problematic as it not only reduces the sample size but also introduces selection bias as poor image quality in UK Biobank is associated with being older, male, non-White, or having higher BMI or blood pressure (Engelmann et al., 2023b). Furthermore, such pipelines can be computationally intensive. Originally, VAMPIRE was semi-automatic and required about four minutes of human time per image (MacGillivray et al., 2015). For fractal dimension, there is now an automated version but this tends to take a similar amount of computer time per image. AutoMorph in my experience is slightly faster than that but still might require a few dozen seconds per image even on a workstation with a graphics processing unit.

My first exposure to this field was through the work of my colleague Ana Villaplana-Velasco who at the time was also a PhD student and was investigating the relationship between cardiovascular disease and fractal dimension. I provided some input on the machine learning methods used for risk prediction. Ana used VAMPIRE to compute retinal fractal dimension in UK Biobank to do her analysis and VAMPIRE proved to be a valuable tool which enabled her analysis (Villaplana-Velasco et al., 2023). However, applying VAMPIRE to close to 100,000 colour fundus images - those that had been found to be of good quality - required a substantial amount of time measured in months. This as well as the high-level of image quality exclusions raised the question whether a more robust and efficient method could be developed.

Traditional pipelines first segment the vessels and then have many additional steps before producing their final output. However, this final output is a single number that deterministically depends on only the image itself. It might be possible to train a machine learning model to directly output the same number, approximately, without intermediary computations. Particularly a deep learning model could be trained to regress the VAMPIRE fractal dimension outputs. While a common narrative is that deep learning models are computationally expensive, in practice this depends on many variables including how large the model is. Furthermore, VAMPIRE already uses a deep learning model for vessel segmentation, and a model that outputs a segmentation generally requires

far more compute than a model that outputs only a single number. Thus, this approach directly approximating the fractal dimension value with a deep learning model not only avoids the subsequent computations but might even be more efficient than the vessel segmentation step itself. This is the motivation for the work presented in this chapter.

3.2 Paper

Robust and Efficient Computation of Retinal Fractal Dimension Through Deep Approximation

Justin Engelmann[1(✉)], Ana Villaplana-Velasco[2], Amos Storkey[3], and Miguel O. Bernabeu[2]

[1] CDT Biomedical AI, School of Informatics, University of Edinburgh, Edinburgh, Scotland
justin.engelmann@ed.ac.uk
[2] Centre for Medical Informatics, University of Edinburgh, Edinburgh, Scotland
[3] School of Informatics, University of Edinburgh, Edinburgh, Scotland

Abstract. A retinal trait, or phenotype, summarises a specific aspect of a retinal image in a single number. This can then be used for further analyses, e.g. with statistical methods. However, reducing an aspect of a complex image to a single, meaningful number is challenging. Thus, methods for calculating retinal traits tend to be complex, multi-step pipelines that can only be applied to high quality images. This means that researchers often have to discard substantial portions of the available data. We hypothesise that such pipelines can be approximated with a single, simpler step that can be made robust to common quality issues. We propose Deep Approximation of Retinal Traits (DART) where a deep neural network is used predict the output of an existing pipeline on high quality images from synthetically degraded versions of these images. We demonstrate DART on retinal Fractal Dimension (FD) - a measure of vascular complexity - calculated by VAMPIRE, using retinal images from UK Biobank that previous work identified as high quality. Our method shows very high agreement with $\text{FD}^{\text{VAMPIRE}}$ on unseen test images (Pearson $r = 0.9572$). Even when those images are severely degraded, DART can still recover an FD estimate that shows good agreement with $\text{FD}^{\text{VAMPIRE}}$ obtained from the original images (Pearson $r = 0.8817$). This suggests that our method could enable researchers to discard fewer images in the future. Our method can compute FD for over 1,000 img/s using a single GPU. We consider these to be very encouraging initial results and hope to develop this approach into a useful tool for retinal analysis. Code for running DART with the trained model is available on GitHub.

Keywords: Retinal fractal dimension · Deep approximation of retinal traits · Robust retinal image analysis

A. Storkey and M. O. Bernabeu—Equal supervision.

B. Antony et al. (Eds.): OMIA 2022, LNCS 13576, pp. 84–93, 2022.
https://doi.org/10.1007/978-3-031-16525-2_9

1 Introduction

Retinal fundus images are non-invasive and low-cost. They are important for ophthalmology and also capture a detailed picture of the retinal vasculature. Thus, they can be used for studying and potentially predicting diseases such as diabetes, stroke, hypertension and neurovascular disease [10]. To analyse the relationships between aspects of the retina and other quantities of interest, retinal traits (also called features, parameters or phenotypes) are used as a quantitative description of a specific aspect of the retinal image. Reducing a complex image to a single, meaningful number is necessary to use standard statistical methods yet a challenging task. It is challenging to identify a potentially salient aspect of the retina in the first place and to then design a method that can reliably quantify this aspect. This is further complicated by the large variability in retinal images stemming from idiosyncrasies of the imaged retinas (e.g. due to retinal diseases or rare phenotypes) and image quality (e.g. due to operator inexperience or time pressures in large scale cohort studies). Thus, pipelines for extracting such retinal traits tend to be complex and comprise of multiple steps, and can only be applied to images of sufficient quality.

Poor image quality is a key problem in retinal image analysis. Particularly for large scale studies such as UK Biobank, many images are of poor quality being blurred, obscured, or hazy [9]. Imaging artefacts such as noise, non-uniform illumination or blur can also lead to poor vessel segmentations [12]. Previous work analysing 2,690 UK Biobank participants found that only 60% had an image that could be adequately analysed by VAMPIRE [9]. Two recent large-scale studies using retinal Fractal Dimension (FD) for predicting cardiovascular disease risk discarded 26% [21] and 43% [16] of the images in UK Biobank. Although necessary, this is unfortunate as it leads to lower sample sizes and makes it hard to study rare diseases in particular.

We hypothesise that it is possible to approximate pipelines for calculating retinal traits with a single, simpler step and propose Deep Approximation of Retinal Traits (DART). Figure 1 gives a high-level overview of our approach. DART trains a deep neural network (DNN) to predict the output of an original method (OM) for calculating a retinal trait. We can then train the model to be robust to image quality issues by synthetically degrading the input images during training and asking the DNN model to predict the output of the OM on the original high quality image. The intuition behind this approach is that obtaining a high quality segmentation of the entire retina is a much harder task than describing an aspect of the vasculature like vascular complexity directly. DART offers a segmentation-free way of computing retinal traits related to the vasculature, but can also be applied to any other retinal image analysis method like feature extraction for disease grading or pathology segmentation.

In the present work, we focus on retinal FD, a key retinal trait that has been used to predict cardiovascular disease risk [16,21] and is associated with neurodegeneration and stroke [6]. FD is a mathematical measure of the complexity of a self-similar object. Applied to the retinal vasculature, FD captures how complex and branching it is which in turn might be a proxy for how healthy the

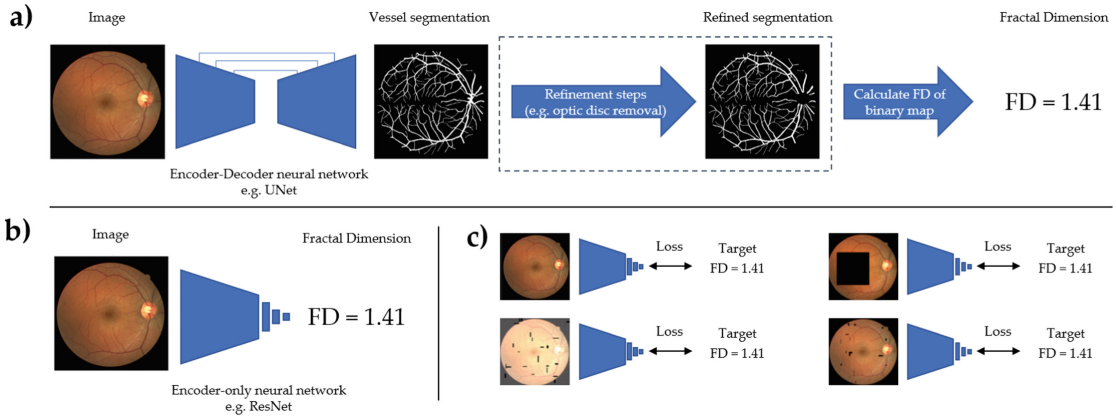


Fig. 1. Overview of our proposed framework. a) A typical pipeline for computing FD: an encoder-decoder neural network for segmentation, potentially some refinement steps like optic disc segmentation and removal, and a method to calculate FD of the segmentation (e.g. box counting or multifractal). b) DART, our proposed approach outputs a deep approximation of FD in a single step using an encoder-only neural network, with drastically reduced complexity. c) We can train our model to be robust to image quality issues by synthetically degrading input images and training our model to minimise the loss between its output and the FD obtained with the original high quality image.

vasculature is. We use FD as calculated by VAMPIRE [15] with the multifractal [14] method as the OM we apply DART to. At minimum, FD^{DART} should have very high agreement with $\text{FD}^{\text{VAMPIRE}}$ on high quality images so that it can be interpreted in the same way. To be a useful method, it should further be robust to image quality issues and efficient. Robustness would enable researchers to discard fewer images than currently necessary while efficiency allows to conduct analyses at large scale without requiring large compute resources.

2 Deep Approximation of Retinal Traits (DART)

2.1 Motivation and Theory

We hypothesise that it is possible to approximate the entire pipeline of an original method (OM) for calculating a retinal trait in a single, simpler step. We denote the distribution of high quality retinal fundus images as X^{HQ}, where each image x_i has dimensions height H, width W, and channels C. The OM can be interpreted as a function f that maps from the image space to one-dimensional retinal trait space (in our case, FD) $f : \mathbb{R}^{HxWxC} \rightarrow \mathbb{R}^1$, i.e. given an image $x_i \in X^{HQ}$ the FD computed by the OM is $FD^{OM} = f(x_i)$. Our goal is to find an alternative function $g : \mathbb{R}^{HxWxC} \rightarrow \mathbb{R}^1$ that is both simpler than f and has high agreement with f for all images of sufficient quality that the OM can be used, i.e. for all $x_i \in X^{HQ}$ $f(x_i) \approx g(x_i)$.

Designing such a simpler function by hand would be very challenging. Thus, we use a deep neural network (DNN). DNNs are universal function approximators in theory and very effective for image analysis in practice. We can then find

a good approximation of f by simply updating the model parameters θ (weights, biases, normalisation layer parameters) to minimise some differentiable measure of divergence between $f(x_i)$ and $g(x_i)$, e.g. mean squared error.

Accuracy. The output of the OM is fully determined by the given image, so we would expect that very high accuracy can be achieved. This contrasts with other problems, e.g. clincians take into account additional information like symptoms and family history, and might disagree with each other or even themselves if shown the same image multiple times.

Simplicity and Efficiency. Some readers might not perceive DNNs as simple or efficient. However, modern pipelines for retinal image analysis tend to use DNNs for vessel segmentation, so not requiring additional steps implies strictly lower complexity both computationally and in terms of required code. Furthermore, segmentation models tend to have an encoder-decoder structure (e.g. UNet) whereas models for classification/regression only need an encoder and small prediction head, making them more parameter-, memory-, and compute-efficient. Finally, given the widespread adoption of deep learning, the frameworks are very mature and can be very efficiently GPU-accelerated.

Robustness. We hypothesise that there images of lower quality that are such that a) current pipelines would not produce a useful FD number, but b) there is still sufficient information to give an accurate estimate of the FD number we would have obtained on a counterfactual high quality image. For example, in an image with an obstruction, only part of the retina might be visible. Thus, the resulting vessel segmentation map would be poor and the FD of this map would be very different from that of the counterfactual high quality image, yet the visible parts of the retina might contain sufficient information about the vascular complexity of the retina as a whole to recover an accurate estimate of the FD.

As we do not observe counterfactual high quality images or objective ground truth FD values, we artificially degrade high quality images with a degradation function $\text{degrade}(x_i) = x_i^{\text{degarded}}$ and train our model to minimise the difference between the predicted FD for the degraded image and the OM's FD for the high quality image $g_\theta(x_i^{\text{degarded}}) \approx f(x_i)$. If there indeed is sufficient information in the degraded images, then our model should be able to predict the OM's FD from the high quality image reasonably well. However, this is a much harder task than matching the OM on high quality images, as the degradations lose information and for a given degraded image there are multiple possible counterfactual high quality images.

2.2 Implementation

Model and Training. Our model consists of a pretrained ResNet18 [4] backbone that extracts a feature map from the images, followed by spatial average

Table 1. Severity levels for the degradations. Brightness, contrast and gamma changes are independently sampled from the given interval. Dimensions in pixels.

Severity	1	2	3	4	5
Brightness/Contrast/ Gamma	±5%	±10%	±15%	±20%	±25%
Mini Artifacts (holes, height, width)	2–20/1–3/5–8	2–24/1–5/5–12	2–28/1–5/5–16	2–32/1–3/5–20	2–40/1–3/5–24
Square Artifacts (side length)	25	50	75	100	125
Chop Artifacts (% of image removed)	10–15	10–25	10–35	10–45	10–50
Advanced Blur (kernel size, sigma)	3–5/0.2–0.5	3–7/0.2–0.7	3–9/0.2–0.8	3–11/0.2–0.9	3–13/0.2–1.0
Gaussian Noise (variance)	1–10	5–10	5–20	5–25	5–30

pool and a small multi-layer perceptron with a two hidden layers with 128 and 32 units, and a single output. Each hidden layer is followed by a layernorm [1] and GELU [5] activation. No activation is applied to the final output. ResNet is a well-established architecture that has been shown to perform competitively with more recent architectures when using modern training techniques [2,19]. We use Resnet18 as it is the most light-weight member of the Resnet family. We initialise the backbones with pre-trained weights on natural images from Instagram [20]. Those images are very different from retinal images, thus this is merely a minor refinement on random initialisation. We resize images to 224×224 pixels for computational efficiency and lower memory requirements. Apart from standard normalisation using channel-wise ImageNet mean and standard deviations, no further preprocessing is done and all 3 colour channels are kept.

We train our model using a batchsize of 256 to minimise the mean squared error between prediction and target after normalizing the target to zero mean and unit variance, using mean and standard deviation from the training data to avoid data leakage. The model output can then be mapped back to FD range by applying the inverse transformation. We use the AdamW optimiser [8] ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay of 10^{-6}) and a cosine learning rate schedule [7]. We train for 35 epochs with a linear learning rate warmup from $\eta_{min} = 10^{-5}$ to $\eta_{max} = 10^{-3}$ for 5 epochs, followed by 3 cycles of 10 epochs each. During each cycle, the current epoch learning rate is set according to a cosine schedule, and after each cycle η_{max} is decayed by taking the square root. We apply generic data augmentations (horizontal ($p = 0.5$) and vertical flip ($p = 0.1$), mild affine transformations ($p = 0.15$, rotation by up to $\pm 10°$, shear of up to $\pm 5°$, and scaling by $\pm 5\%$)) as well as the image degradations described in the next section with $p = 0.75$ (sampling all 5 levels uniformly) to the images during training. We used Python 3.9 with PyTorch and timm [18]. Our code for running DART, including the trained model, is available here: https://github.com/justinengelmann/DART_retinal_fractal_dimension.

Synthetic Degradations. We focus on three types of quality issues in retinal images [9,12]: Lighting issues, artifacts/obstructions, and imaging issues. To simulate general lighting issues, we independently change brightness, contrast and gamma of the image. To simulate artifacts/obstructions and severely inconsistent lighting, we introduce one of three artifacts: 1) many smaller rectangular holes placed across the retina, b) a single large square hole, or c) we "chop" off the bottom or top part of the image. The latter is inspired by the observation that in UK Biobank some images only have the top or bottom part properly illuminated. To simulate general imaging issues, we add pixel-wise Gaussian noise and blur the image. Standard isotropic Gaussian blur kernels do not mimic realistic image blur, so we use an advanced anisotropic blurring technique developed for image super-resolution [17] where the standard deviations for both dimensions of the kernel are sampled independently, and the kernel is then rotated and has some noise added before being applied to the image. These synthetic degradations are inspired by common retinal imaging quality issues but do not perfectly mirror them. Our goal here is to test the feasibility of using DART to recover good FD estimates from severely degraded images. Thus, our degradations heavily feature artifcats and blur, both of which remove information from the images. If DART can recover good FD estimates under these challenging conditions, then this would be reason to think that it will also work under more realistic, yet less challenging conditions.

We specify degradation parameters for five levels of severity, shown in Table 1. For a given level, we sample parameters for each image independently from the given ranges. Degradations are applied after images have already been downsized to 224×224. We apply an artifact with $p = 0.2 * s$ where s is the severity. If an image was chosen to have an artifact applied to it, we then choose Mini Artifacts with $p = 0.85$, Square Artifact with $p = 0.10$, and Chop Artifact with $p = 0.05$. Degradations are implemented using the albumentations package [3].

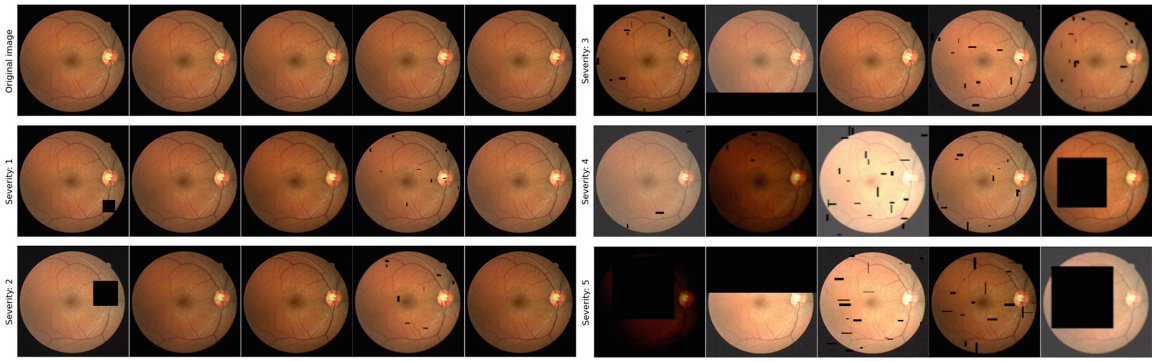


Fig. 2. Random examples of synthetically degraded versions of the same fundus image. Best viewed zoomed in, especially for the advanced blur. UK Biobank asks to only reproduce imaging data where necessary, so we demonstrate the degradations on an image taken from DRIVE [13] which is similar in appearance to those in UK Biobank.

Table 2. Agreement between $FD^{VAMPIRE}$ obtained on high quality images, and FD^{DART} for different levels of degradation measured on 14,907 held-out test set images.

Degradations	R^2	Pearson r (p-value)	Spearman r (p-value)	OLS Regression fit
None	0.9160	0.9572 (0.0000)	0.9561 (0.0000)	y = 0.01 + 1.00x
Severity 1	0.8957	0.9467 (0.0000)	0.9446 (0.0000)	y = 0.01 + 0.99x
Severity 2	0.8859	0.9414 (0.0000)	0.9396 (0.0000)	y = 0.01 + 0.99x
Severity 3	0.8623	0.9287 (0.0000)	0.9282 (0.0000)	y = 0.00 + 1.00x
Severity 4	0.8309	0.9116 (0.0000)	0.9103 (0.0000)	y = 0.01 + 0.99x
Severity 5	0.7773	0.8817 (0.0000)	0.8840 (0.0000)	y = 0.02 + 0.99x

3 Experiments

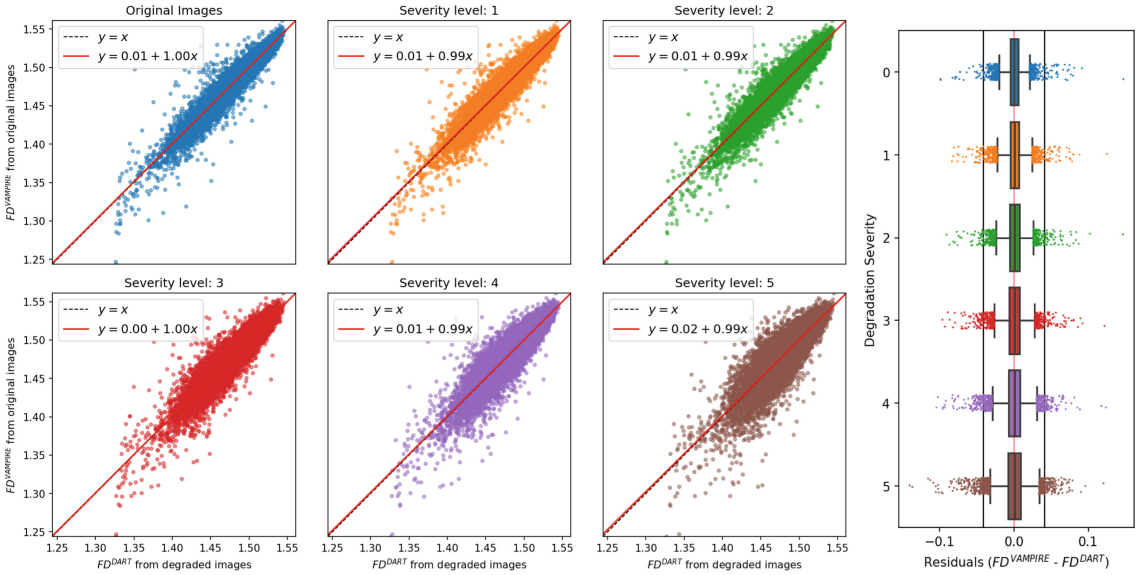
3.1 Data

We apply our DART framework multi-fractal FD [14] calculated with VAMPIRE [15]. We use only images from UK Biobank that had been identified as high quality (top 60% of in terms of quality) in a previous study that used FD for cardiovascular disease risk prediction [16]. Thus, for those images $FD^{VAMPIRE}$ should be reliable and can be considered as a reasonable "ground-truth". We randomly split the data into train, validation, and test sets containing 70, 10, and 20% of the participants in UK Biobank, resulting in 52,242/7,478/14,907 images belonging to 32,300/4,614/9,229 participants in each set. We split at the participant level such that no images of the same participant occur in different sets. Images are cropped to square to remove black non-retinal regions and processed at 224×224 as described above.

3.2 Results

Agreement and Robustness. We find very high agreement between $FD^{VAMPIRE}$ and FD^{DART} on the original images with Pearson $r = 0.9572$ and $r^2 = 0.9160$. Table 2 shows results for different levels of degradations. When degrading the images and asking our model to predict the $FD^{VAMPIRE}$ obtained from the high quality image, agreements goes down as the images become more degraded, which is what we would expect as these degradations remove substantial information about the retinal vasculature. However, despite this, we still observe good agreement with the $FD^{VAMPIRE}$ obtained on the original image even at severity level 5 where extreme degradations are applied (Pearson $r = 0.8817$ and $R^2 = 0.7773$). This suggests that DART can recover good estimates of the retinal trait that would have been obtained from a counterfactual high quality image even if the available image has very poor quality. Thus, this might allow for discarding much fewer images than currently necessary.

For comparison, a previous study comparing FD for arteries and veins separately between VAMPIRE and SIVA [11] found very poor agreement between the measures of the two tools ($R^2 = 0.139$ and $R^2 = 0.168$ for arteries and



(a) Scatterplots of $\mathrm{FD}^{\mathrm{DART}}$ against $\mathrm{FD}^{\mathrm{VAMPIRE}}$ obtained from original images for different levels of degradation.

(b) Boxplots of the residuals.

Fig. 3. Agreement results for 14,907 held-out test set images. Best viewed zoomed in. **a)** Red line: best linear fit; dashed black line: $y = x$. **b)** Faint red line: $x = 0$; vertical black lines: $\pm$ one interquartile range (IQR) of $\mathrm{FD}^{\mathrm{VAMPIRE}}$ for reference. (Color figure online)

veins, respectively). Another study comparing vessel caliber-related retinal traits obtained with VAMPIRE, SIVA, and IVAN found that they agreed with Pearson rs of 0.29 to 0.86. Thus, the observed agreement between $\mathrm{FD}^{\mathrm{VAMPIRE}}$ and $\mathrm{FD}^{\mathrm{DART}}$ with a Pearson $r = 0.9572$ and $R^2 = 0.9160$ is very high, and even when DART is applied the most degraded images the agreement (Pearson $r = 0.8817$ and $R^2 = 0.7773$) is higher than what could be expected when using two different tools on the same high quality images.

Finally, our method shows very low bias even as degradation severity is increased (Fig. 3). The best OLS fit is very close to the identity line for all levels of severity, or equivalently, the optimal linear translation function from $\mathrm{FD}^{\mathrm{DART}}$ to $\mathrm{FD}^{\mathrm{VAMPIRE}}$ is almost simply the identity function. This also implies that no post-hoc adjustment for image quality is needed and $\mathrm{FD}^{\mathrm{DART}}$ values obtained for images of varying quality are on the same scale out-of-the-box. As degradation severity increases, the variance of the residuals also increases but most residuals are still less than one interquartile range (IQR), a robust equivalent of the standard deviation, even when applying the strongest degradation.

Speed. Images were loaded into RAM so that hard disk speed is not a factor. We then measured the time it took to process all 52,242 training images, including normalisation, moving them from RAM to GPU VRAM, as well as the time to move the results back to RAM. We used a modern workstation (Intel i9-9920X 24 core CPU, single Nvidia RTX A6000 24GB GPU, 126GB of RAM) and a

batchsize of 440. With ResNet18 as backbone, our model processed all 52,242 images in 48.5s ± 93.6 ms (mean ± std over 5 runs), yielding a rate of 1,077 img/s.

4 Conclusion

We have shown that we can use DART to approximate the multi-step pipeline for obtaining $FD^{VAMPIRE}$ with very high agreement. Our resulting model can compute FD^{DART} for over 1,000 img/s using a GPU. Furthermore, our model can compute FD^{DART} values from severely degraded images that still match the $FD^{VAMPIRE}$ values obtained on the high quality images well. This could allow researchers interested in studying retinal traits to discard fewer images than currently necessary and thus have higher sample sizes. We consider these to be very encouraging initial results.

There are a number of directions for future work. First, the proposed framework can be easily applied to other retinal traits like vessel tortuosity or width, or FD as calculated by other pipelines. We would expect that this would be similarly successful. Second, the robustness of the resulting DART model should be evaluated in more depth and the cases with extreme residuals should be manually examined. We expect that robustness can be further improved, especially if we identify common failure cases and use those as data augmentations. Third, many straight-forward, incremental technical improvements should be possible such as improved training procedures to further increase performance, trying different architectures and resolutions, and speeding up inference speed further through common tricks like fusing batch norm layers into the convolutional layers. Finally, we hope that our approach will eventually enable other researchers to conduct better analyses, e.g. by not having to discard as many images and thus having a larger sample size available.

Acknowledgements. We thank our friends and colleagues for their help and support. This research has been conducted using the UK Biobank Resource (project number 72144). This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. Support from Diabetes UK (grant 20/0006221) and Fight for Sight (grant 5137/5138) is gratefully acknowledged.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bello, I., et al.: Revisiting ResNets: improved training and scaling strategies. arXiv preprint arXiv:2103.07579 (2021)
3. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020). https://doi.org/10.3390/info11020125

4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)
6. Lemmens, S., Devulder, A., Van Keer, K., Bierkens, J., De Boever, P., Stalmans, I.: Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: assessment of a potential biomarker. Front. Neurosci. **14**, 16 (2020)
7. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
9. MacGillivray, T.J., et al.: Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. PLoS ONE **10**(5), e0127914 (2015)
10. MacGillivray, T., Trucco, E., Cameron, J., Dhillon, B., Houston, J., Van Beek, E.: Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. Br. J. Radiol. **87**(1040), 20130832 (2014)
11. McGrory, S., et al.: Towards standardization of quantitative retinal vascular parameters: comparison of SIVA and VAMPIRE measurements in the Lothian Birth Cohort 1936. Transl. Vis. Sci. Technol. **7**(2), 12 (2018)
12. Mookiah, M.R.K., et al.: A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. Med. Image Anal. **68**, 101905 (2021)
13. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
14. Stosic, T., Stosic, B.D.: Multifractal analysis of human retinal vessels. IEEE Trans. Med. Imaging **25**(8), 1101–1107 (2006)
15. Trucco, E., et al.: Novel VAMPIRE algorithms for quantitative analysis of the retinal vasculature. In: 2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC), pp. 1–4. IEEE (2013)
16. Velasco, A.V., et al.: Decreased retinal vascular complexity is an early biomarker of MI supported by a shared genetic control. medRxiv (2021)
17. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914 (2021)
18. Wightman, R.: PyTorch image models (2019). https://doi.org/10.5281/zenodo.4414861
19. Wightman, R., Touvron, H., Jégou, H.: ResNet strikes back: an improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
20. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
21. Zekavat, S.M., et al.: Deep learning of the retina enables phenome-and genome-wide analyses of the microvasculature. Circulation **145**(2), 134–150 (2022)

3.3 Conclusion

The results in this work are very encouraging, the agreement between DART and VAMPIRE when both receive high quality images exceeded what I expected to be possible. Likewise, even when receiving severely degraded images as input, DART was able to match the output VAMPIRE gave for the original, non-degraded version remarkably well. This suggests that DART could be more robust which might mean that it requires fewer image quality exclusions and offers better signal-to-noise ratio. The computational efficiency of DART is very promising, as it allows large datasets to be analysed even on lower end hardware. Pipelines like VAMPIRE require dozens of seconds or even minutes per image, so DART is a few orders of magnitude faster. Thus, for example, processing all colour fundus images in UK Biobank would previously have taken weeks or months, and now could be done in less than a day. Of course, DART stands on the shoulder of giants and requires an existing pipeline to approximate, so it is an approach that is complementary with current approaches for retinal image analysis.

Future work should apply DART to real-world data and examine its repeatability and robustness in more detail. Indeed, this is what I am doing in Chapter 3 and Chapter 4, respectively, but additional validation would always be helpful. Furthermore, DART could be extended and improved in a number of ways. First, using higher quality "groundtruths" for training might improve its repeatability. I manually inspected cases where DART and VAMPIRE disagreed most on the held-out data, and it appeared that most of these were poor quality images that had not been filtered out (data not shown). For instance, one image in the validation set had a VAMPIRE fractal dimension of 1.25, far below the next lowest values (see for example Figure 3 of the paper), which appeared to be an outlier due to poor image quality rather than genuinely extremely low vessel complexity. Filtering out these outliers might increase the robustness of a future DART model, as we do not want to model to learn to replicate errors in the original pipeline. Similarly, we could try to remove noise from the output of the original pipeline that we use for training by applying that pipeline multiple times to each image. Each time, we make small modifications to the image such as horizontal flipping or minor changes in brightness. Averaging the output across the minor changes would then cancel out random noise and give a more consistent ground truth for training. Second, the training procedure for DART could be improved to explicitly encourage repeatability. Concretely, we could use multiple augmented versions of each image in a training batch, and penalise the model not just for disagreement with the value

obtained from the original pipeline, but also for the variance across its outputs for the different versions of the same image. Third, DART could be extended to other retinal traits such as vessel density or tortuosity. Indeed, these three potential improvements are already the subject of on-going work.

Chapter 4

Application of DART to real-world clinical data

4.1 Introduction

"Oculomics", research relating retinal traits and systemic health (Ify Mordi and Emanuele Trucco, 2022; Trucco et al., 2013; Wagner et al., 2022), is commonly focused on datasets that are not representative of the wider population, such as UK Biobank (Chua et al., 2019) which was collected for research purposes and does not perfectly represent the UK population (Fry et al., 2017), or AlzEye which only includes patients who attended Moorfields Eye Hospital (Wagner et al., 2022). Both of these dataset further do not include younger adults, with subjects being at least 40 years old (Chua et al., 2019; Wagner et al., 2022). This raises the question whether oculomics research could transfer to real-world primary care settings with a mixed-age adult population. This is interesting from a research perspective, e.g. whether associations between retinal traits and systemic health are limited to older adults, and for potential practical applications as retinal image-based risk prediction of systemic conditions - if one day it becomes practically useful - would be particularly interesting in primary care settings.

For the work presented in this chapter, we had access to a small dataset of colour fundus images and record cards from a university-based optometry clinic at Glasgow Caledonian University, which is reflective of a primary-care setting and included patients between 18 and 81 years of age. This makes it an interesting test bed, although it has several limitations. Most notably, the relatively small size of 96 individuals and that the only source of information about prevalent systemic health conditions was a

field related to “General Health” on the record cards. A secondary motivation of this work was to apply and validate DART. If DART did not generalise well beyond UK Biobank and failed to produce meaningful outputs, then it would be unlikely that there would be any statistically significant associations with systemic health.

4.2 Paper

tvst

Artificial Intelligence

Retinal Fractal Dimension Is a Potential Biomarker for Systemic Health—Evidence From a Mixed-Age, Primary-Care Population

Justin Engelmann[1,2], Stephanie Kearney[3], Alice McTrusty[4], Greta McKinlay[3], Miguel O. Bernabeu[1,5,*], and Niall Strang[3,*]

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK
[2] School of Informatics, University of Edinburgh, Edinburgh, UK
[3] Department of Vision Sciences, Glasgow Caledonian University, Glasgow, UK
[4] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[5] The Bayes Centre, University of Edinburgh, Edinburgh, UK

Correspondence: Justin Engelmann, Centre for Medical Informatics, Usher Institute, University of Edinburgh, NINE Edinburgh BioQuarter, Plot 9 Little France Road, Edinburgh EH16 4UX, UK. e-mail: justin.engelmann@ed.ac.uk

Received: September 5, 2023
Accepted: March 3, 2024
Published: April 12, 2024

Keywords: fractal dimension; oculomics; retinal image analysis

Citation: Engelmann J, Kearney S, McTrusty A, McKinlay G, Bernabeu MO, Strang N. Retinal fractal dimension is a potential biomarker for systemic health—evidence from a mixed-age, primary-care population. Transl Vis Sci Technol. 2024;13(4):19, https://doi.org/10.1167/tvst.13.4.19

Purpose: To investigate whether fractal dimension (FD), a retinal trait relating to vascular complexity and a potential "oculomics" biomarker for systemic disease, is applicable to a mixed-age, primary-care population.

Methods: We used cross-sectional data (96 individuals; 183 eyes; ages 18–81 years) from a university-based optometry clinic in Glasgow, Scotland, to study the association between FD and systemic health. We computed FD from color fundus images using Deep Approximation of Retinal Traits (DART), an artificial intelligence–based method designed to be more robust to poor image quality.

Results: Despite DART being designed to be more robust, a significant association ($P < 0.001$) between image quality and FD remained. Consistent with previous literature, age was associated with lower FD ($P < 0.001$ univariate and when adjusting for image quality). However, FD variance was higher in older patients, and some patients over 60 had FD comparable to those of patients in their 20s. Prevalent systemic conditions were significantly ($P = 0.037$) associated with lower FD when adjusting for image quality and age.

Conclusions: Our work suggests that FD as a biomarker for systemic health extends to mixed-age, primary-care populations. FD decreases with age but might not substantially decrease in everyone. This should be further investigated using longitudinal data. Finally, image quality was associated with FD, but it is unclear whether this finding is measurement error caused by image quality or confounded by age and health. Future work should investigate this to clarify whether adjusting for image quality is appropriate.

Translational Relevance: FD could potentially be used in regular screening settings, but questions around image quality remain.

Introduction

Retinal color fundus imaging rapidly and non-invasively captures a detailed picture of the retinal vasculature while being widely available and low cost. Thus, retinal image–derived traits are being investigated as biomarkers for systemic health, a field also known as "oculomics."[1–3] Fractal dimension (FD), a retinal trait relating to the complexity of the retinal vasculature, has emerged as a particularly promising candidate that is associated with cardiovascular[4,5] and neurovascular[6,7] disease. FD is calculated from vessel segmentations and captures how complex the branching structure of the blood vessels is. A lower FD indicates a less complex vasculature, which might be an indicator for poorer retinal vessel health. This in turn could indicate poorer vessel health systemically, and,

indeed, individuals with lower FD are at higher risk of cardiovascular disease,[4] for example. Thus, retinal FD derived from color fundus images could serve as a potential biomarker for systemic health in research and clinical practice that could identify individuals at high risk who could then be examined in more detail.

However, research to date has focused on cohort studies such as the UK Biobank or data from secondary care settings such as AlzEye.[8,9] These are generally older and not representative of the broader population. Mean age is 57.8 ± 8.6 years in the UK Biobank[10] and 68.4 ± 13.9 years in AlzEye. In addition to the participants being older, the UK Biobank has more female participants and fewer who are socioeconomically deprived than the general UK population,[11] and AlzEye participation was predicated on hospital attendance. This leaves open the question of whether FD as a biomarker for systemic health is applicable to mixed-age, primary-care populations—populations that span from young adults all the way to advanced age who have been assessed in a standard primary-care setting.

Previous work looking at retinal vascular traits and systemic health is also limited by the exclusion of large amounts of data due to image quality, on the order of 20% to 45% even for datasets such as the UK Biobank that are specifically collected for research.[2,4,5] Even rejecting one in five images would drastically limit the utility of FD in clinical practice. Furthermore, older, less healthy, male, and non-White subjects are at higher risk of being excluded due to image quality.[12] Thus, quality-based exclusions introduce selection bias that could exacerbate existing disparities in healthcare research.

Recently, a novel artificial intelligence–based method for computing FD has been proposed: Deep Approximation of Retinal Traits (DART).[13] Traditional methods such as VAMPIRE[14] calculate FD from binary segmentations of the blood vessels, and even small imperfections in those segmentations can affect the calculations, resulting in a high bar for minimum image quality. DART uses a deep learning model that was trained to output the same FD as VAMPIRE but during training received original high-quality images and degraded versions of those, such as with altered brightness or contrast to simulate lighting issues, blur, or simulated artifacts. The model was tasked to output the number VAMPIRE gave for the original high-quality image and thus learned to ignore variations in image quality and take into account all available information about vessel structure to estimate FD. Intuitively, even if only part of the vessels is highly visible in a fundus image, the image still contains substantial information about the vasculature. DART can leverage this information, whereas traditional approaches are not robust to such cases. DART has shown very high internal validity on held-out UK Biobank images, matching VAMPIRE with a Pearson correlation of 0.9572 when DART received the original image and a correlation of 0.8817 when DART received severely degraded images instead (both $P < 0.0001$).

In this work, we investigated the potential of using FD as a potential biomarker for systemic health in clinical practice by studying a mixed-age, primary-care population. To avoid introducing bias and to evaluate the robustness of FD under challenging, real-world conditions, we made no image quality exclusions.

Methods

Glasgow Caledonian University Cohort

Clinical records from eye examination appointments that took place between 2017 and 2022 at the Vision Centre at Glasgow Caledonian University (GCU), Glasgow, Scotland, United Kingdom, were analyzed. As part of the eye examinations, a record card was completed, and color fundus images were captured for each patient. The study was undertaken in accordance with the tenets of the Declaration of Helsinki. Ethical approval was obtained from the GCU School of Health Sciences Ethical Committee prior to the commencement of data collection (HLS/LS/A22/003). Participants provided written consent to have their anonymized clinical records used for research purposes. Retinal images were obtained using swept-source optical computed tomography (DRI OCT Triton Plus; Topcon, Tokyo, Japan).

The record cards were either digitized if they were in physical paper form or extracted from the patient management system. All of the images were available digitally on the device and exported in JPG format at a resolution of 2000 × 1312 pixels or greater and linked to the record cards. The following information was then collated in a consistent format: age at visit, sex, and information about the general health status from the "history and symptoms" field in the record card.

We had access to 183 images linked to 96 records belonging to 58 female and 38 male patients. In nine cases, only a single image was available on the device, presumably because only one eye was imaged during the examination. For consistency, we used the most recent visit for each individual, and all available images were included in our analysis. The data extraction process was labor intensive; thus, we had access to only

Table 1. Population Characteristics Stratified by Systemic Health Status

	All	Prevalent Systemic Condition	No Systemic Condition
n	96	46	50
Female, %	60.4	52.2	68.0
Age (y), median ± IQR	61.80 ± 33.20	62.79 ± 9.60	55.28 ± 39.51

one visit per patient. During data extraction, we prioritized extracting data for as many patients as possible at the cost of not extracting longitudinal data. The median age at visit was 61.80 years (interquartile range [IQR], 33.20) with the youngest and oldest patients being 18 and 81 years, respectively. Notably, 22 patients were under the age of 30, an age group that is not available in UK Biobank or AlzEye where the youngest subjects are 37[10] and 40,[9] respectively.

Systemic Health Information

We used information about systemic health from the "history and symptoms" field of the record cards to analyze the relationship between systemic health and FD. This information was recorded in the context of an optometric examination and thus was generally coarse grained with varying levels of detail, primarily consisting of very short descriptions and commonly used clinical abbreviations (e.g., "Good, no problems," "diabetes", "HBP"). We stratified individuals into two groups based on this information: those described as having any systemic health condition (i.e., any non-ocular health condition) and those with no mention of any such condition. Table 1 gives an overview of the two groups.

Although the level of detail for the systemic health information is limited at times, this approach should have high positive predictive value; that is, individuals in the "prevalent systemic conditions" group would be very likely to actually have systemic conditions. Sensitivity, on the other hand, would be imperfect. Prevalent conditions might not have been mentioned by the patient, might not have been deemed sufficiently relevant to record, or might have been undiagnosed at the time of visit. This should lead to lower apparent effect sizes. Another limitation is that we do not have information about the severity or duration of the conditions. This should likewise lead to lower apparent effect sizes than if we could account for severity. Thus, we expect that the limitations of this variable biased our analysis to be more conservative, and any apparent differences between the two groups are likely to be true differences but with underestimated effect sizes; in other words, the risk of a type 1 error was low and that of a type 2 error was high.

Computing FD and Image Quality Annotation

To compute the FD of the images, we used DART,[13] which is based on the multifractal FD of VAMPIRE.[2,14,15] All images could be successfully processed in less than a minute on a consumer-grade, desktop central processing unit (CPU), and no images were excluded from analysis. In addition to computing FD, we manually annotated retinal image quality on a four-level ordinal scale (very good, good, poor, or very poor). The images were annotated by an experienced research optometrist with 10 years of clinical experience (S.K.). Previous work by Laurik-Feuerstein et al.[16] found good intergrader agreement for a similar four-level ordinal taxonomy for color fundus images. Poor-quality fundus images are common even in research datasets such as the UK Biobank, where researchers typically exclude 20% to 40% of the available images. In our dataset, 34.6% of the images were rated as very good, 33.9% as good, 28.1% as poor, and 3.4% as very poor. The proportion of our images rated as poor or very poor (31.5%) is comparable to the 20% to 45% of images that are typically excluded in the UK Biobank.[2,4,5] Examples for each of the four quality levels are shown in Supplementary Figure S1.

Statistical Analysis

We first analyzed the FD values computed by DART to see whether they showed the expected association with age and whether there was an association with image quality or sex that we needed to adjust for. We then used a linear mixed-effects model at the eye level with FD as the dependent variable and a random intercept per patient to adjust for the two eyes of an individual not being independent. The data was analyzed using the statsmodels[17] package (version 0.13.5) in Python 3.9.13. We used a threshold of $P < 0.05$ for statistical significance throughout.

For retinal traits relating to vessel caliber, differences in magnification due to refractive error (RE) can change the apparent size of vessels.[18,19] We expect that for FD this is not the case, as FD relates to the branching structure of the vessels and especially

because VAMPIRE, on which DART is based, uses skeletonized vessel maps and should thus be invariant to apparent caliber. However, we also empirically investigated this to ensure that our decision not to include RE did not affect our results.

Results

FD, Image Quality, Age, and Sex

The mean FD in our data was 1.4904, with a standard deviation (SD) of 0.0391. An increase in image quality issues by one level was associated with a decrease in FD by 0.026 (95% confidence interval [CI], 0.031–0.021) in absolute terms, or by 0.656 SD (95% CI, 0.788–0.524; $P < 0.001$). This suggests that, although DART is designed to be more robust to image quality issues, there might still be an effect that must be adjusted for.

Figure 1 shows scatterplots of FD versus age for the raw data, as well as when FD was being adjusted for image quality in a mixed-effects model. Increasing age was associated with a significant decrease in FD. In a univariable model, an additional decade of age was associated with a decrease in FD by 0.232 SD (95% CI, 0.323–0.141; $P < 0.001$). When adjusting for image quality issues, this changed to a decrease by 0.172 SD per decade (95% CI, 0.235–0.109; $P < 0.001$), which is consistent in direction and magnitude with what has been reported in the literature.[20] Furthermore, although the effect of age on FD did decrease after adjusting for image quality issues, the direction was the same and the magnitude comparable. This suggests that, although the effect of image quality is significant, it does not preclude discovering meaningful associations in our data.

In addition to the linear association between age and FD, we also observed that the FD variance was higher in older patients. Some patients over 60 had FD comparable to those of patients in their 20s (Fig. 1a). This suggests that patients could follow characteristically different trajectories, with only some seeing their FD decrease substantially as they age. The findings persisted when adjusting for image quality using the coefficients from the linear mixed-effects model (Fig. 1b) and when taking the mean of both eyes per patient (Fig. 1c). Sex showed no significant association with FD in a univariable model ($P = 0.252$), when adjusted for image quality ($P = 0.156$) and when adjusted for image quality and age ($P = 0.377$).

FD and Systemic Disease

As our preliminary analysis suggested that age and image quality are significantly associated with FD, we adjusted our model for these variables. We fit the following mixed-effects model with a random intercept per patient: FD ~ age + image quality + prevalent systemic conditions + (1|patient). Table 2 shows the resulting model. Prevalent systemic conditions were significantly associated with a decrease in FD by 0.246 SD (95% CI, −0.477 to −0.015; $P = 0.037$). In a univariable model, the decrease was 0.461 SD (95% CI, −0.833 to −0.088; $P = 0.015$) and was 0.361 SD (95% CI, −0.612 to −0.110; $P = 0.005$) when just adjusting for image quality only.

FD and Refractive Error

We split eyes into three groups based on their RE: hyperopic (RE > 1.5; 30 eyes), myopic (RE < −1.5; 50 eyes), and emmetropic. When added to the

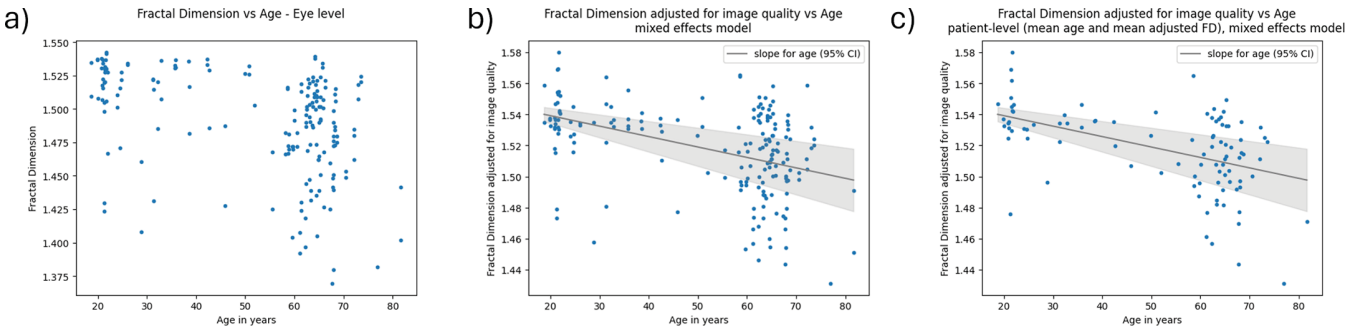


Figure 1. Association of retinal FD and age. (**a**) Raw data. (**b**) FD adjusted for image quality using a mixed-effects model with random intercepts for each patient at the eye-level (i.e., each point corresponds to one eye). (**c**) The same model as in part (**b**) but with values at the patient level (i.e., each point is one patient). For parts (**b**) and (**c**), the coefficient for age is −0.007 per decade of age (95% CI, −0.009 to −0.004; $P < 0.001$).

Table 2. Mixed-Effects Model on FD

Variable	β	95% CI	P
Intercept	39.711	39.369 to 40.054	0.000
Age (in decades)	−0.158	−0.221 to −0.095	0.000
Image quality (ordinal, higher is worse)	−0.631	−0.750 to −0.512	0.000
Prevalent systemic conditions	−0.246	−0.477 to −0.015	0.037
Random effect group variance	0.187	—	—

All coefficients are in standard deviation units of FD.

model in Table 2, neither being hyperopic nor being myopic had a significant effect ($P = 0.159$ and $P = 0.204$, respectively), and prevalent systemic conditions had the same coefficient ($\beta = -0.246$) with a slightly reduced P value of 0.041, possibly due to residual cofounding between RE and age. When looking at RE adjusted only for age, there likewise were no significant associations with FD, with P values of 0.159 and 0.204 for being hyperopic and myopic, respectively. Thus, RE does not appear to have influenced the FD measurements in our dataset and would not have meaningfully affected our results if we had included it in our model.

Discussion

FD as a Biomarker for Systemic Health

We observed a significant association between prevalent systemic conditions and FD, even when controlling for age and image quality. This confirms previous findings that FD might be a biomarker for systemic health. This is further corroborated by the observation that FD had higher variance in older patients. Importantly, our data was collected during clinical practice in a primary-care setting and included patients from 18 to 81 years of age, and no images were excluded from analysis based on quality. Thus, our work suggests that FD can be used in an even more challenging setting that is closer to clinical reality than what previous work considered.

A major limitation is the coarseness of the information about prevalent systemic conditions and the lack of information about incident systemic conditions. However, as we argued in the Methods section, we expect that the variable for prevalent systemic conditions will have high positive predictive value; thus, the observed effect is likely to be a true effect and the effect size will be underestimated. Future work should further explore FD as a biomarker for systemic health in primary-care data, ideally with linkage to other medical records for more detailed information about systemic health.

Should Image Quality be Adjusted for?

We adjusted for image quality due to its strong association with FD, which is commonly done in the literature.[4] The underlying assumption is that image quality issues affect the computation of FD itself and thus must be adjusted for to recover retinal vascular complexity. This notion initially appears convincing, but upon reflection is potentially problematic.

Retinal vascular complexity is supposed to be a proxy for systemic health; however, age and poorer health have been found to be associated with not just vascular complexity but also image quality itself.[12] Plausibly, frailer individuals could be more difficult to image (e.g., because of difficulty sitting still). Furthermore, age-related changes to the eye, such as miosis and cataract, also decrease expected image quality.

Figure 2 shows a simplified causal diagram for FD measurements, with a potential direct effect of age on retinal vascular complexity (dashed pink arrow). To our knowledge, this effect has not been conclusively established in the literature but could plausibly exist. The dashed orange arrow indicates a potential effect of image quality on FD, which is the reason why image quality is commonly adjusted for.

However, observe that, even if we had a fully robust method where this arrow would not be present, we would expect to observe a significant association between FD and image quality, confounded by systemic health. In that case, adjusting for image quality would be inappropriate and would bias the association between FD and systemic health toward the null. Even if our method is not perfectly robust and there is an effect of image quality on FD (i.e., poor-quality images lead not only to measurement noise but also to a systematic bias in FD), then controlling for image quality could likewise bias our analysis.

Thus, we argue that adjusting for image quality is potentially problematic and should be more critically

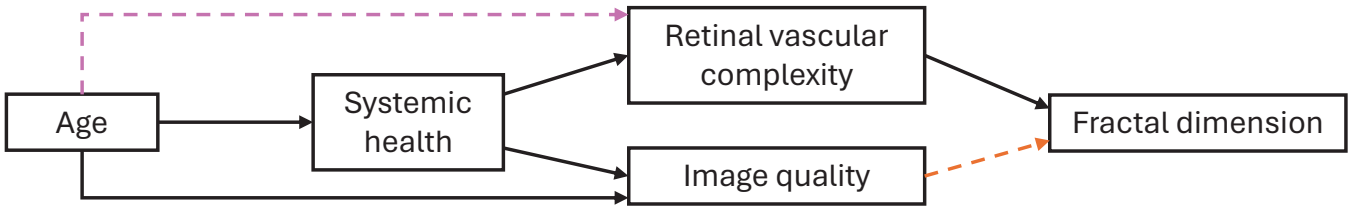


Figure 2. Simplified causal diagram for FD measurements. The *dashed orange arrow* indicates the undesirable potential effect of image quality on FD. The *dashed pink arrow* indicates a potential direct effect of age on retinal vascular complexity that could occur in healthy aging.

considered in oculomics research. This issue further highlights the need for robust methods for computing retinal traits, as more robust methods would reduce the need to adjust for image quality in the first place.

Note that this issue cannot be side-stepped by only examining high-quality images and excluding the rest. As mentioned in the Introduction, these exclusions introduce selection bias and exacerbate inequalities in healthcare research,[12] in addition to reducing sample sizes and statistical power. Future work should look at the repeatability and robustness to image quality of DART and traditional approaches to add empirical evidence to the question of whether adjusting for image quality can be avoided.

Age and FD

Our finding that some older patients have FD similar to those of younger patients suggests that FD might not substantially decrease with age in all individuals. Likewise, the increased variance of FD in older patients suggests that individual trajectories might be quite heterogeneous. It also suggests that FD does not merely change with age, lending credence to the hypothesis that it captures meaningful biological changes and that it could provide information about systemic health beyond what age itself provides. However, our analysis here is cross-sectional and future work should investigate this in longitudinal data.

Conclusions

In addition to the aforementioned limitation relating to the systemic health information, our work has several additional limitations. First, only a single quality annotator was used, although they were very experienced, and comparable quality taxonomies for color fundus imaging have good repeatability according to the literature. Future work ideally should use fully automated methods such as the recently proposed QuickQual[21] that avoid introducing subjectivity and allow better comparison of quality annotations across different works. Second, although DART is more robust to image quality than traditional approaches, some images might still be too poor in quality. In the present dataset, even the "very bad" images had at least some visible vasculature that could provide sufficient information for a reasonable FD estimate (see Supplementary Fig. S1). Nevertheless, this issue should be further investigated in future work. Third, despite the promising results in the present work and literature at large, future work should also more closely investigate what specific vascular changes are captured by FD.

In summary, we found that prevalent systemic health conditions are associated with a significant decrease in FD in a mixed-age, primary-care population. Furthermore, although FD generally decreases with age, our data suggest that FD might not substantially decrease in everyone. This heterogeneity could be due to systemic health, further supporting FD as a potential biomarker for systemic health. Future work should study the relationship between FD and age in longitudinal data and clarify whether adjusting for image quality is appropriate.

Acknowledgments

Supported by a grant from UK Research and Innovation (EP/S02431X/1 to JE) as part of the Centre of Doctoral Training in Biomedical AI at the School of Informatics, University of Edinburgh; Fondation Leducq Transatlantic Network of Excellence (17 CVD 03 to MOB); a grant from the Engineering and Physical Sciences Research Council (EP/X025705/1); British Heart Foundation and The Alan Turing Institute Cardiovascular Data Science Award (C-10180357); and Diabetes UK (20/0006221). MOB and NS acknowledge funding from Fight for Sight (5137/5138); SCONe projects funded by the Chief Scientist Office, Edinburgh & Lothians Health Foundation, Sight Scotland, and the Royal College of Surgeons of Edinburgh; RS Macdonald Charitable Trust.

Disclosure: **J. Engelmann**, None; **S. Kearney**, None; **A. McTrusty**, None; **G. McKinlay**, None; **M.O. Bernabeu**, None; **N. Strang**, None

* MOB and NS contributed equally to this article.

References

1. Wagner SK, Fu DJ, Faes L, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol*. 2020;9(2):6.
2. MacGillivray TJ, Cameron JR, Zhang Q, et al. Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS One*. 2015;10(5): e0127914.
3. Mordi I, Trucco E. The eyes as a window to the heart: looking beyond the horizon. *Br J Ophthalmol*. 2022;106(12):1627–1628.
4. Villaplana-Velasco A, Engelmann J, Rawlik K, et al. Decreased retinal vascular complexity is an early biomarker of MI supported by a shared genetic control. *medRxiv*. 2021;12:16.2126 7446.
5. Zekavat SM, Raghu VK, Trinder M, et al. Deep learning of the retina enables phenome- and genome-wide analyses of the microvasculature. *Circulation*. 2022;145(2):134–150.
6. Lemmens S, Devulder A, Van Keer K, Bierkens J, De Boever P, Stalmans I. Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: assessment of a potential biomarker. *Front Neurosci*. 2020; 14:16.
7. McGrory S, Ballerini L, Doubal FN, et al. Retinal microvasculature and cerebral small vessel disease in the Lothian Birth Cohort 1936 and Mild Stroke Study. *Sci Rep*. 2019;9(1):6320.
8. Luben R, Wagner S, Struyven R, et al. Retinal fractal dimension in prevalent dementia: the AlzEye Study. *Invest Ophthalmol Vis Sci*. 2022;63(7):4440–F0119.
9. Wagner SK, Hughes F, Cortina-Borja M, et al. AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ Open*. 2022;12(3):e058552.
10. UK Biobank. Data-Field 21003: age when attended assessment centre. Available at: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21003. Accessed April 4, 2024.
11. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034.
12. Engelmann J, Storkey A, LLinares MB. Exclusion of poor quality fundus images biases health research linking retinal traits and systemic health. *Invest Ophthalmol Vis Sci*. 2023;64(8):2922.
13. Engelmann J, Villaplana-Velasco A, Storkey A, Bernabeu MO. Robust and efficient computation of retinal fractal dimension through deep approximation. In: *OMIA: I*nternational Workshop on Ophthalmic Medical Image Analysis. Cham: Springer Nature: 2022:84–93.
14. Trucco E, Ballerini L, Relan D, et al. Novel VAMPIRE algorithms for quantitative analysis of the retinal vasculature. In: *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2013:1–4.
15. Stosic T, Stosic BD. Multifractal analysis of human retinal vessels. *IEEE Trans Med Imaging*. 2006;25(8):1101–1107.
16. Laurik-Feuerstein KL, Sapahia R, Cabrera DeBuc D, Somfai GM. The assessment of fundus image quality labeling reliability among graders with different backgrounds. *PLoS One*. 2022;17(7):e0271156.
17. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. 2010.
18. Wong TY, Wang JJ, Rochtchina E, Klein R, Mitchell P. Does refractive error influence the association of blood pressure and retinal vessel diameters? The Blue Mountains Eye Study. *Am J Ophthalmol*. 2004;137(6):1050–1055.
19. Bengtsson B. The variation and covariation of cup and disc diameters. *Acta Ophthalmol*. 1976;54(6):804–818.
20. Cheung CY, Thomas GN, Tay W, et al. Retinal vascular fractal dimension and its relationship with cardiovascular and ocular risk factors. *Am J Ophthalmol*. 2012;154(4):663–674.e1.
21. Engelmann J, Storkey A, Bernabeu MO. QuickQual: lightweight, convenient retinal image quality scoring with off-the-shelf pretrained models. In: Antony B, Chen H, Fang H, Fu H, Lee CS, Zheng Y, eds. *Ophthalmic Medical Image Analysis*. Cham: Springer Nature, 2023:32–41.

Supplementary Figure 1: Two random examples for each of the four ordinal levels of quality, as graded by SK.

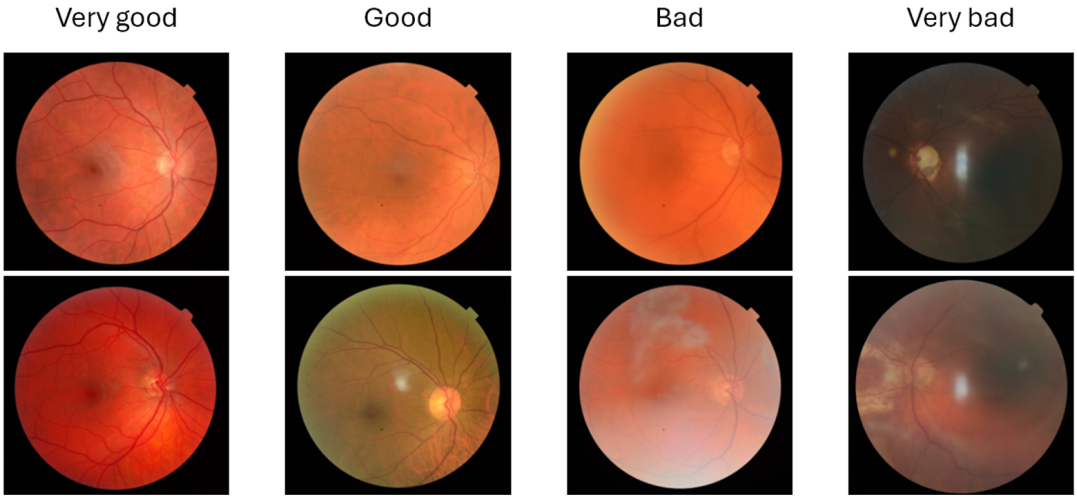


Figure 4.1: Supplementary Figure.

4.3 Conclusion

The results here are encouraging as they serve as initial validation of DART - if it did not capture anything meaningful or did not generalise beyond UK Biobank, it is unlikely we would have observed significant associations with age and systemic health - and suggests that fractal dimension computed with DART might be applicable to real-world datasets, which are less carefully quality controlled and potentially more representative than research datasets like UK Biobank. However, there are many limitations of this work, most importantly the small size of the dataset and the information about prevalent systemic conditions. The latter was a crude binary label and based on free text information collected during an optometry visit. As discussed in the paper, we would expect this label to have high positive predictive value but low negative predictive value. These limitations are reason to think the analysis is more likely to be biased towards the null, so these findings are indeed encouraging. Still, ultimately we need to be cautious in interpreting these results and validation in larger datasets with higher quality disease information is needed.

Although a minor part of the analysis, in the paper we also found that there was not apparent association between refractive error and fractal dimension, suggesting that the magnification effect of refractive error does not appear to affect the computation of fractal dimension. This is a useful result as other retinal traits such as vessel calibre had been found to be affected refractive error and ideally need to be adjusted for it (Wong et al., 2004). Knowing that fractal dimension does not need to be similarly adjusted is useful and means that it can be applied in datasets where information about refractive error is not available. Of course, it would be ideal to repeat this analysis in a larger dataset to ensure that this is the case. I analysed this in a second dataset where there likewise was no significant association (data not shown) and if there was a particularly large effect we likely would have detected it, but until we look at a more comprehensive dataset we cannot rule out a small effect of refractive error. In the future, I would like to examine the relationship between refractive error and various retinal traits in more detail.

While we adjusted for image quality to be more conservative, the causal relationship between age, systemic health, retinal traits and retinal image quality should be explored in more detail. Recalling that image quality itself appears to be associated with key patient characteristics including age, blood pressure and body mass index (Engelmann et al., 2023b), even if the computation of our retinal trait of interest was not affected by

image quality at all, we would expect to find a correlation between quality and the retinal trait. This is provided the retinal trait is itself correlated with age and health-related attributes, but that is likely the case as most retinal traits are studied as potential correlates of systemic health, which in turn correlates with age and health-related attributes. Thus, adjusting for image quality in our analyses might introduce bias towards the null. I think this is a subtle, yet important question that will be difficult to answer conclusively. Examining the repeatability and robustness of retinal traits like fractal dimension, as is done in Chapter 4, is a small step towards resolving that problem, as the lower the influence of image quality on the retinal trait, the lower the need to adjust for image quality. Ultimately, we would need to have a better understanding of the determinants of image quality which will require substantial and careful research.

While providing some initial evidence that associations between retinal traits and systemic health might generalise from research datasets to real-world clinical data, the present analysis was only statistical and limited to looking for an association. To be clinically useful, retinal traits like fractal dimension would need to provide a substantial improvement in predictive power over easily available risk factors such as age, sex, smoking status, blood pressure, or body mass index. This is a much higher bar than an association with a p-value below 0.05. So while these are encouraging results, I think it is likely that we are still a ways away from clinical utility.

Chapter 5

Repeatability and robustness of DART compared with a pipeline following the traditional paradigm for computing fractal dimension

5.1 Introduction

Ideally, tools for computing retinal traits studied in relation to systemic health should be highly repeatable across multiple images of the same eye during the same visit. For instance, fractal dimension is studied as a measure of vessel complexity which in turn is a proxy for vascular health. It would be somewhat implausible for vascular health to change substantially during a visit, so fractal dimension, too, should not vary substantially across multiple images during the same visit. Low repeatability implies poor signal-to-noise ratio. Note that technically high repeatability is only a necessary condition for a retinal image analysis tool being useful, but not sufficient. In theory, a measure could be highly repeatable but not capture anything biologically meaningful. As an extreme case, suppose we had a tool that produced an entirely arbitrary number for each individual, but reliably produced the same number for each individual each time without noise. Such a tool would be perfectly repeatable, yet useless for statistical analyses. However, in practice high repeatability is at least good reason to think that a tool is likely capturing something meaningful.

Retinal image analysis tools should also be robust to quality issues, since that increases sample sizes available for research, reduces the magnitude of selection bias due to quality exclusions (Engelmann et al., 2023b), and is likely to lead to increased repeatability, too. For potential clinical adoption of retinal image analysis tools, robust-

ness is also key. In research, excluding a quarter (Zekavat et al., 2022) or even half (MacGillivray et al., 2015; Villaplana-Velasco et al., 2023) of the data is not uncommon. In clinical practice, working only half the time would greatly limit the applicability of a tool.

Previous work by MacGillivray et al. (2015) looked at how many images in UK Biobank were analysable with VAMPIRE (Trucco et al., 2013), the correlation between left and right eye measurements, as well as the inter-grader agreement as VAMPIRE is semi-automatic and requires manual inputs for each image. However, MacGillivray et al. (2015) did not examine the repeatability of VAMPIRE for repeated images of the same eyes during the same visit. Automated methods such as AutoMorph (Zhou et al., 2022) and DART avoid the subjectivity introduced by manual inputs and in a sense have perfect inter-grader agreement, as two people processing the same image would get the same number. But the question of repeatability across different images of the same eye during the same visit remains, i.e. how much variation in measurements is there due to small variations in the images themselves.

Huang et al. (2016) looked at variation due to different manual graders for vessel segmentation and different automated segmentation methods, but also looked at repeatability across different fundus cameras for 12 subjects and across repeated images for a single subject, finding that there is variation between cameras and between repeated images, respectively. While the study is interesting and impressive regarding the range of sources of variability examined, there are a number of limitations that make the results hard to interpret. The discussion here will focus on some of the major ones and will not be exhaustive. First, for repeatability across repeated imaging, the sample size is quite small. Second, while a number of automated vessel segmentation methods are considered, none of them are deep learning-based. Many retinal image analysis pipelines, including VAMPIRE and AutoMorph now use deep learning-based segmentation methods as they tend to work well, especially compared to previous approaches. Third, repeatability is evaluated using relative standard deviation, i.e. the standard deviation across measurements divided by their average. This metric is sensitive to the location of a measurement, i.e. shifting measurements by a fixed amount changes the relative standard deviation. Especially for an abstract measure like fractal dimension, it is not immediately obvious how it should be interpreted. A key question is how large the variation in fractal dimension is relative to its variation across the population or relative to how much it differs between cases and controls.

An additional motivation for this work is that examining robustness is important to make better recommendations for what threshold to use for image quality exclusions. Ideally, such decisions should be made in a data-driven way, by examining empirically at level level of image quality measurements become unacceptably noisy.

The work presented in this chapter tries to work towards an answer to these questions. By examining both DART and AutoMorph, it further serves to validate DART and to compare the novel, unconventional paradigm that DART uses with a more traditional paradigm for computing fractal dimension, exemplified by AutoMorph. An observant reader might think that examining repeatability and robustness of DART, and comparing it to traditional approaches would have been sensible first steps for validating it, prior to the work presented in the previous chapter. And indeed, I was interested in investigating these questions for the better part of my PhD, ever since I developed DART. However, suitable data with multiple images per eye and visit was hard to come by. I thank my colleague Diana Moukaddem, who collected the Caledonia dataset during her PhD and kindly shared it with me which enabled this analysis.

5.2 Paper

Applicability of Oculomics for Individual Risk Prediction: Repeatability and Robustness of Retinal Fractal Dimension Using DART and AutoMorph

Justin Engelmann,[1,2] Diana Moukaddem,[3] Lucas Gago,[4] Niall Strang,[3] and Miguel O. Bernabeu[1,5]

[1]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[3]Department of Vision Sciences, Glasgow Caledonian University, Glasgow, United Kingdom
[4]Departament de Matemátiques i Informática, Universitat de Barcelona, Barcelona, Spain
[5]Bayes Centre, University of Edinburgh, Edinburgh, United Kingdom

Correspondence: Justin Engelmann, Nine BioQuarter, Plot 9 Little France Rd., Edinburgh EH16 4UX, UK; justin.engelmann@ed.ac.uk.

NS and MOB contributed equally as senior authors.

Received: January 2, 2024
Accepted: May 6, 2024
Published: June 6, 2024

Citation: Engelmann J, Moukaddem D, Gago L, Strang N, Bernabeu MO. Applicability of oculomics for individual risk prediction: Repeatability and robustness of retinal fractal dimension using DART and AutoMorph. *Invest Ophthalmol Vis Sci.* 2024;65(6):10. https://doi.org/10.1167/iovs.65.6.10

Purpose. To investigate whether fractal dimension (FD)–based oculomics could be used for individual risk prediction by evaluating repeatability and robustness.

Methods. We used two datasets: "Caledonia," healthy adults imaged multiple times in quick succession for research (26 subjects, 39 eyes, 377 color fundus images), and GRAPE, glaucoma patients with baseline and follow-up visits (106 subjects, 196 eyes, 392 images). Mean follow-up time was 18.3 months in GRAPE; thus it provides a pessimistic lower bound because vasculature could change. FD was computed with DART and AutoMorph. Image quality was assessed with QuickQual, but no images were initially excluded. Pearson, Spearman, and intraclass correlation (ICC) were used for population-level repeatability. For individual-level repeatability, we introduce measurement noise parameter λ, which is within-eye standard deviation (SD) of FD measurements in units of between-eyes SD.

Results. In Caledonia, ICC was 0.8153 for DART and 0.5779 for AutoMorph, Pearson/Spearman correlation (first and last image) 0.7857/0.7824 for DART, and 0.3933/0.6253 for AutoMorph. In GRAPE, Pearson/Spearman correlation (first and next visit) was 0.7479/0.7474 for DART, and 0.7109/0.7208 for AutoMorph (all $P < 0.0001$). Median λ in Caledonia without exclusions was 3.55% for DART and 12.65% for AutoMorph and improved to up to 1.67% and 6.64% with quality-based exclusions, respectively. Quality exclusions primarily mitigated large outliers. Worst quality in an eye correlated strongly with λ (Pearson 0.5350–0.7550, depending on dataset and method, all $P < 0.0001$).

Conclusions. Repeatability was sufficient for individual-level predictions in heterogeneous populations. DART performed better on all metrics and might be able to detect small, longitudinal changes, highlighting the potential of robust methods.

Keywords: oculomics, fractal dimension, quantitative ophthalmology

Retinal color fundus images are low cost, fast to acquire, and noninvasive; yet they provide a detailed picture of the retinal vasculature. Thus color fundus imaging could provide biomarkers for systemic disease,[1] a field of study sometimes referred to as *oculomics*.[2] A particularly promising candidate biomarker is retinal fractal dimension (FD), which describes the complexity of the vessel structure. A less complex vasculature could indicate poorer retinal vascular health, and this in turn might correlate with vascular health elsewhere in the body. For instance, lower FD is associated with cardiovascular disease outcomes like myocardial infarction[3–5] and has also been studied in relation to neurovascular conditions like dementia.[6,7]

Those are exciting and promising results, but whether they can be translated into useful tools for clinical practice is still an open question. Effect sizes and increases in predictive performance over baselines using basic, easily available information like age, sex, and smoking status are typically small. Thus, for individual level predictions, retinal traits like FD would need to have very low measurement noise, yet this has to date been understudied.

Studies also often exclude a large fraction of the available images due to insufficient quality, on the order of 25% to 45%[3,5,8] in datasets like UK Biobank that were specifically collected for research. These exclusions are especially problematic for clinical applicability of oculomics. If the measurement of the retinal trait of interest (e.g., FD) fails a quarter or half of the time, then that makes it impractical. Furthermore, being older, non-White, or male increases the risk of having poor-quality images,[9] and thus these exclusions introduce

selection bias. This means that results of existing oculomics research might not apply equally well to everyone, and if we wanted to use FD in clinical practice, the measurement would systematically fail more often for some people (e.g. those of non-White ethnicity).

Thus we set out to investigate whether FD-based oculomics could be used for individual risk prediction by first evaluating FD's repeatability at a population and an individual level, without any image quality exclusions. We use two tools for computing FD: AutoMorph,[10] which follows the established paradigm of segmentation, skeletonization, and box counting; and deep approximation of retinal traits (DART),[11] which uses a novel paradigm of directly computing FD via a deep learning model that is trained to be more robust to image quality. We then examine how repeatability changes with the level of image exclusions because of quality and look at the relationship between measurement noise and image quality at the level of individual eyes.

Methods

Datasets

We included two datasets for this study: First, the "Caldonia" dataset, which was collected at Glasgow Caledonian University, Glasgow, Scotland, United Kingdom. Second, the "Glaucoma Real-world Appraisal Progression Ensemble" or "GRAPE" dataset,[12] which was collected at the Eye Center of the Second Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang, China. Both studies had ethical approval and adhered to the Declaration of Helsinki. Participants in both studies signed a written consent form. Table 1 shows a detailed overview of both datasets.

The Caledonia dataset was collected on a Topcon DRI OCT Triton Plus as part of a PhD project looking at choroidal thickness. Thus, the main focus was acquisition of optical coherence tomography (OCT) volume scans, but fortunately color fundus images were acquired at the same time for most scans. Multiple scans were taken on a single day, though in some cases the data collection was repeated due to insufficient OCT quality. Thus five subjects underwent imaging on two different days, three subjects on three days, and one subject on four days. We included every eye with at least five available color fundus images. The subjects were 20 students and six PhD candidates at Glasgow Caledonian University.

The GRAPE dataset was collected on a Topcon TRC-NW8 (108 eyes) and a Canon CR-2 PLUS AF (88 eyes) during clinical practice. The first examination was for suspected glaucoma, with subsequent follow-up visits to monitor progression. Subjects were treated with IOP decreasing drugs after their first visit and only those with glaucoma are included in the study. We included all eyes that had a baseline and follow-up color fundus image, taking follow-up images from the first follow-up visit that had an available image.

We analyze both datasets to examine FD in 132 subjects, imaged at two different locations with three different devices, covering a large age range, different ethnicities and both healthy and glaucomatous eyes. The Caledonia dataset provides relatively ideal conditions for repeatability, namely many images per eye, collected on the same or a handful of days in a research setting, in young adults that are generally easier to image. However, the color fundus images were not a focus during the data collection, so the quality will likely vary at least somewhat.

The GRAPE dataset is a longitudinal dataset with only one image per eye per visit and a mean follow-up time of 18.3 months. FD a measure of retinal vascular complexity and general vascular health, which could conceivably change between visits. Thus, even a perfectly repeatable method would not be expected to produce the same measurement for both visits. Furthermore, data were collected during clinical practice in a population that included over 60 year olds and thus image quality is likely more mixed. Especially because FD is calculated from the vasculature but for glaucoma the optic disc is most important, so images that were sufficient for the clinical purposes during collection might be suboptimal for calculating FD.

Based on these considerations, we expect Caledonia to provide a slightly optimistic estimate for repeatability, whereas GRAPE should provide a pessimistic, lower-bound for repeatability. Taken together, these two datasets will allow us to characterize the repeatability of FD well.

Computation of FD

We used DART (short for "deep approximation of retinal traits")[11] and AutoMorph[10] to calculate FD from the color fundus images. AutoMorph is a multistep pipeline consisting of a deep learning model for vessel segmentation followed by skeletonization and the box counting method to compute

Table 1. Overview of the Datasets Used, Reporting Statistics for All Subjects We Included

	Caledonia	GRAPE	Combined
Subjects	26	106	132
Eyes	39	196	235
Color fundus images	377	392	769
Interval Mean ± SD [Min–Max]	2.4 ± 6.6 months [12 hours–26 months]	18.3 ± 13.3 months [5.3–53.1]	/
Female sex	13 (50%)	52 (49%)	65 (49%)
Age in years Mean ± SD [Min–Max]	24.0 ± 3.6 [18–33]	41.7 ± 15.0 [18–74]	38 ± 15 [18–74]
Ethnicity	17 White, 6 Asian, 2 Black, 1 Middle Eastern	Presumably primarily or entirely Chinese	Primarily Chinese, 17 White, 6 Asian, 2 Black, 1 Middle Eastern
Ocular health	3 cases of amblyopia, otherwise healthy	All glaucoma (103 [97%] open angle; 3 [3%] angle closure)	106 glaucoma, 3 amblyopia, 23 healthy
Refractive status	12 hyperopes, 7 myopes, 7 emmetropes	Not available, but presumably mostly myopes as most subjects have open- angle glaucoma	Primarily myopes, at least 7 emmetropes, at least 12 hyperopes

Interval is the time between baseline and follow-up image in GRAPE, and between first and last image in Caledonia.

FD. This is a similar approach to other tools for calculating FD like VAMPIRE.[13] Changes to the AutoMorph pipeline (e.g., varying the box sizes used for the FD calculation) might affect its repeatability. Our goal in the present article is not to propose a new algorithm for computing FD or analyze potential modifications that could be made to AutoMorph but simply to use it as provided. We want to analyze the repeatability of AutoMorph as it is released, and this matches what the vast majority of researchers would do in practice, especially those without extensive programming knowledge.

DART, on the other hand, uses a single deep learning model to directly output FD from the image. DART's deep learning model was trained to replicate the output of VAMPIRE on images from UK Biobank with sufficient quality to apply VAMPIRE and achieved very high internal validity (Pearson correlation of 0.9572 on 14,907 held-out validation images). DART was trained to not just replicate VAMPIRE's output but also to be more robust to image quality. During the training progress, the model either received the original, high quality image or a poor quality version of it obtained by randomly adjusting brightness, contrast, and gamma, simulating imaging problems with anisotropic blur and gaussian noise, and adding artefacts to the images. These might ordinarily affect the output of pipelines to compute FD, which is undesirable. However, whether DART received the original or a degraded version of it, it was tasked to output the FD VAMPIRE calculated from the high quality image either way, encouraging it to ignore variations in image quality and thus be more robust.

We chose these methods because they are both openly available on Github, allowing researchers to easily and freely access them without seeking prior permission. Furthermore, AutoMorph is a method following the traditional paradigm of segmentation, skeletonization, and box counting, whereas DART uses a novel yet less tried paradigm. For transparency, we want to make the reader aware that two authors of this work (JE and MB) were involved in the development of DART and thus—despite our best efforts to be neutral and objective—the reader should critically examine the present work.

Previous work comparing retinal traits computed with different tools found poor to moderate interchangeability.[14,15] We think that the interchangeability of DART and AutoMorph, while tangential to our main research question, might be of interest to the reader. We conducted this analysis retrospectively and use the quality exclusion threshold of QuickQual $P(\text{bad}) < 0.8$ that we later recommend based on our results. See the next section for a description of QuickQual. We use mean values per eye to reduce measurement noise and find that in Caledonia, DART and AutoMorph agree with a Pearson correlation of 0.6390 and Spearman correlation of 0.7096 (both $P < 0.0001$). For GRAPE, they agreed with a Pearson correlation of 0.4418 and Spearman correlation of 0.4914 (both $P < 0.0001$). Bland-Altman plots are shown in Supplementary Figure S1. Thus both tools show a level of interchangeability that is comparable to what previous work reported for other tools.

Assessment of Image Quality

We assessed image quality with QuickQual, a recently proposed very efficient method that leverages vector embeddings from a foundation model for natural images and obtains state-of-the-art on the EyeQ dataset. Concretely, we use the "MEga Minified Estimator" (MEME) version of QuickQual that provides a continuous "good-bad" quality score (probability of being bad, $P(\text{bad})$ for short) as opposed to EyeQ's original three-way classification into "good," "useable," and "bad" images. Images in the "useable" class were mapped to $P(\text{bad}) = 0.5$ during the training of QuickQual-MEME so that the model learns to put imperfect yet useable images in-between good and bad images. Thus QuickQual-MEME's quality score is ideal for examining different levels of quality exclusions and ranges from $P(\text{bad}) = 0$ indicating very good images to $p(\text{bad}) = 1$ indicating very bad images.

Statistical Analysis

Terminology. There are many terms relating to measurement noise (e.g., "agreement," "reliability," "reproducibility," and "repeatability") that are used differently by different authors.[16] In many investigations, a key focus is to compare measurements that are made by different human annotators, something that is not applicable here as we use deterministic, fully-automatic methods. For simplicity, we use the term "repeatability" throughout while carefully explaining what data we analyzed (above) and what metrics we calculate (below).

Population-Level Metrics. For population-level metrics, we use Pearson and Spearman rank correlation, as well as the intraclass correlation coefficient (ICC). Pearson correlation is a linear measure and sensitive to outliers. Spearman rank correlation is a robust, nonparametric measure that uses the Pearson correlation of the ranks of both variables, instead of the raw values of the variables themselves. Thus Spearman correlation is robust to outliers and captures how similar the ranking is across both sets of measurements. Pearson and Spearman are applicable to paired measurements (e.g., two FD values of the same eye).

For more than two measurements per eye, as we have in the Caledonia dataset, we need to use the ICC instead. Though commonly referred to as *the* ICC, there are multiple versions, some of which are to examine agreement between different raters, which is not applicable here. We use the ICC as described by Bartlett and Frost,[16] namely $ICC = \frac{(SD\ of\ subject's\ true\ values)^2}{(SD\ of\ subject's\ true\ values)^2 + (SD\ measurement\ error)^2}$, where SD is standard deviation. In words, this captures how much of the variation in the data is due to between-subject variation as a fraction of the total variation of the data. The total variation in the denominator is composed of the between-subject variation and variation due to measurement error. If measurement error was 0, the ICC would be 1. As measurement error becomes large relative to between-subject variation, the ICC decreases and approaches 0 in the limit.

The ICC is an abstract, unobservable quantity that we need to estimate based on the data. We estimate the subject's true values by taking the mean of all available images of a given eye, and then take the SD of all eyes. The SD measurement error is equivalent to the within-subject SD. Which population we choose to calculate the inter-subject variation is a key design choice that must be made taking into account the specific context, which is also stressed by Bartlett and Frost.[16] As we examine the measurement error in the Caledonia dataset, estimating the between-subject variation in the same data is the natural choice. However, as that dataset consists of healthy adults, we would expect their vascular health to be good and thus FD to have little variation. Therefore we present the reader with two versions of the ICC. One

Investigative Ophthalmology & Visual Science

using the inter-subject variation from Caledonia ("ICC"), and one adjusted version using the inter-subject variation from the combined Caledonia and GRAPE datasets ("Adj. ICC").

Finally, we also report Pearson and Spearman in the Caledonia dataset to enable easier comparison with the measures in GRAPE and because we expect that more readers are familiar with those measures. Because we have more than two images per eye, we take two approaches to calculate Pearson and Spearman. First, we take the first and last available image per eye to make an objective, yet arbitrary choice. Second, we randomly sample one pairing per eye, calculate the correlations, and then repeat this process 20,000 times, reporting median values with an empirical 95% confidence interval. Note that bootstrapped confidence intervals for Pearson correlation can have inaccurate coverage.[17] Our sampling-based approach is similar and thus the confidence interval for Pearson might not be reliable.

Individual-Level Metrics. The metrics above summarize repeatability in a population. However, we are also interested in repeatability at an individual level. Thus we propose the relative SD λ as a metric of individual-level measurement noise, $\lambda = \frac{SD\ of\ FD\ within\ eye}{SD\ of\ FD\ across\ eyes}$. λ expresses how large the variation of FD within an eye is compared to the variation of FD between eyes. As SD is a sum of squared mean deviations, large errors are weighted more heavily, which we think is desirable in this context. Conceptually it is similar to Pearson correlation and ICC, although for λ smaller values are better. A λ of 0 implies no measurement noise, and the larger λ gets, the more noise there is. For convenience, we express λ in %. For λ, we use the SD of FD across eyes as estimated from the combined dataset, for the reasons explained in the previous section.

Robustness to Image Quality. To examine the relationship between repeatability and image quality and to evaluate the robustness of the two methods, we first look at how λ changes in Caledonia as we exclude a larger share of images due to image quality. We consider exclusion percentages from 0% to 50%, which was chosen because it covers and spans slightly beyond typical values in the oculomics literature. Next, we relate λ and the worst image quality in a given eye. We take the worst rather than the mean quality as a single outlier could lead to a high λ. Recall that SD is based on squared differences from the mean, and thus a single large deviation influences the SD more than many small deviations. We compute the Pearson correlation between λ and worst image quality, and further plot them against each other to examine the relationship between the two. This could give some insight into whether there is a critical level of quality where repeatability decreases quickly. Finally, QuickQual-MEME's quality score is the probability of an image being bad. However, probabilities are constrained quantities which can be an issue for Pearson correlation. Thus we also evaluate the Pearson correlation between λ and the raw logit value logit(*P*(bad)) (i.e., the raw output of QuickQual-MEME) before applying the logistic linkage function. Note that the logistic linkage function is a monotonic, and thus the Spearman correlation is the same in both cases.

Results

Fractal Dimension and Population-Level Repeatability

The SD of DART FD at the eye-level was 0.00733 in Caledonia, 0.03653 in GRAPE, and 0.03557 in the combined dataset. For AutoMorph, the SDs were 0.02421, 0.08926, and 0.08841, respectively. Table 2 shows different population-level metrics for both methods and datasets.

Individual-Level Repeatability

Figure 1 shows the distributions of individual-level measurement noise λ for both datasets and methods. λ is generally higher for the GRAPE dataset, which is what we expected based on the dataset characteristics. In the Caledonia dataset the third quantile, or seventy-fifth percentile, is less than 10% for DART and less than 20% for AutoMorph. For the GRAPE dataset, the third quantiles are less than 35% for both methods.

In Caledonia, the worst values of λ were 81.02% and 199.96% for DART and AutoMorph, respectively. The image with the worst value was the same for both methods. The best values were 0.40% and 5.07%. Figure 2 shows the best and worst examples. For the worst example, the high SD within the eye is driven by two very badly illuminated images. Removing those would change λ to 2.71% for DART and to 8.19% for AutoMorph. For the best example for DART, DART gave virtually the same FD for all images while AutoMorph had a bit more variation. On the other hand, for the best example for AutoMorph, AutoMorph has slightly less but similar variation to DART.

Figure 2 also conveys a sense of how λ relates to where a subject might be placed in the distribution of FD across subjects. A λ of ∼5% as in the rightmost subplot is not ideal, but ultimately places the individual in a similar location. For example, for AutoMorph all values are around the third quartile and for DART between the median and third quartile. However, it would preclude us from detecting differences

Table 2. Population-Level Metrics of Repeatability for Both Methods and Datasets Without Quality Exclusions, Including all Available Images

	Caledonia						GRAPE	
	First and Last Image of Each Eye		20,000 Random Pairs Per Eye		All Images		First and Next Visit	
	Pearson	Spearman	Pearson	Spearman	ICC	Adj. ICC	Pearson	Spearman
DART	0.7857 (0.6252–0.8825)***	0.7824 (0.6199–0.8805)***	0.6845 (0.1876–0.9483)	0.7561 (0.5836–0.8893)	0.8153	0.9907	0.7479 (0.6789–0.8038)***	0.7474 (0.6783–0.8034)***
AutoMorph	0.3933 (0.0888–0.6306)***	0.6253 (0.3859–0.7858)***	0.3235 (−0.0683–0.8676)	0.6097 (0.3472-0.8043)	0.5779	0.9494	0.7109 (0.6340–0.7740)***	0.7208 (0.6459–0.7819)***

Adj. ICC, adjusted ICC; ICC, intraclass correlation coefficient.
Higher is better for all measures. 95% confidence intervals in brackets.
*** denotes $P < 0.0001$. Note that the coverage of the Pearson correlation confidence interval for the 20,000 random pairs might have inaccurate coverage.

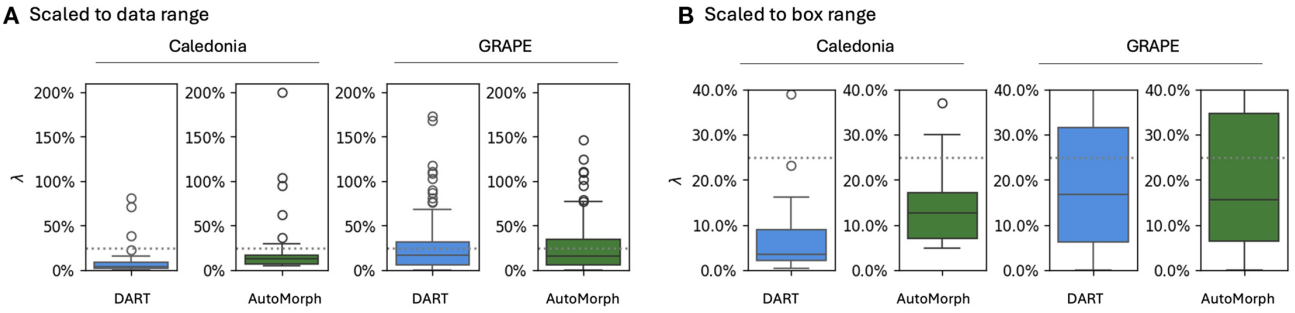


Figure 1. Boxplots for individual-level measurement noise λ for both datasets and methods without quality exclusions, including all available images. The y-axis has the same scale for all boxplots in a given subplot. Subplot **A)** is scaled to the data range, subplot **B)** is scaled such the boxes themselves fit. The horizontal dashed line indicates λ = 25% as a visual aid.

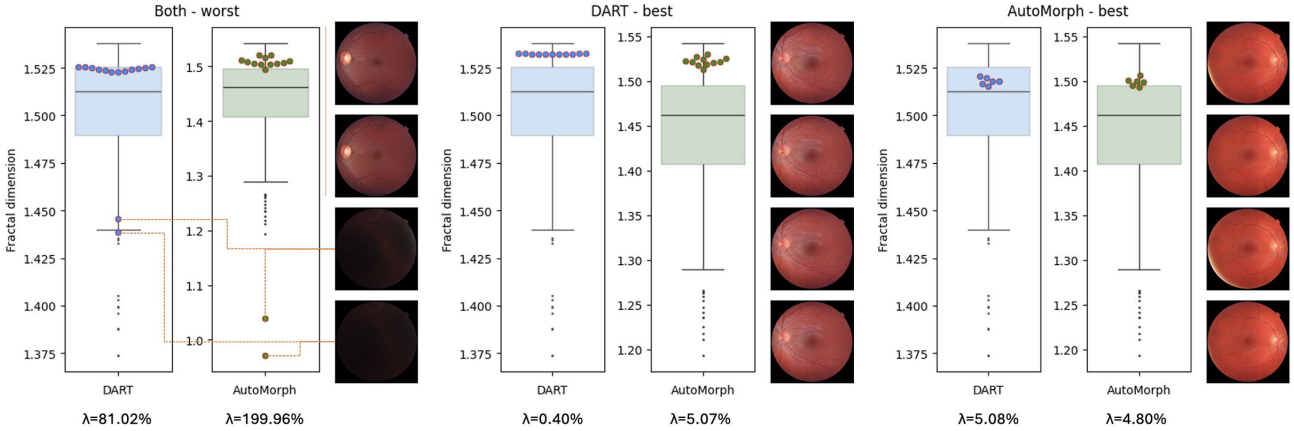


Figure 2. Eyes with the worst and best λ for both methods. The pale boxplots in the background show the distribution of mean FD per eye across both datasets for visual reference. The colored points with red rim show individual FD measurements for the given eye. For the worst example, we show the images for the two outliers at the bottom. All other images are randomly sampled. Below the boxplots, we indicate λ for the eye. Note that the scale of the boxplot for AutoMorph is different for the worst example as it was rescaled to fit the two outliers.

that are around 5% of the SD of FD or smaller (e.g., in longitudinal images).

Robustness to Image Quality

Figure 3 shows how the distribution of λ changes as images are excluded because of quality. When no images are excluded, the highest λ for DART and AutoMorph are 81.02% and 199.96%, respectively. This decreases to 16.19% and 37.00% when the worst 5% of images are excluded. The median λ for DART is 3.55% without any exclusions, which gradually decreases to 1.67% as more images are excluded. For AutoMorph, the median is 12.65% without exclusions which decreases to 6.77% with increasing levels of exclusions.

Interestingly, the minimum λ for AutoMorph was 4.80% without exclusions, and still 3.08% with 35% of the images being excluded. This contrasts with DART, which had a constant minimum λ of 0.40% even without exclusions. Thus, AutoMorph's best case λ was 7.5 to 12 times higher than that of DART. AutoMorph's median was 3.5 times higher without exclusion and three times higher at best, namely when 40% of the images were excluded. Overall, exclusion of poor quality images primarily removes very large outliers, whereas median and best case repeatability only change slightly.

The Pearson correlation between λ and worst quality in a pair for DART was 0.7550 in Caledonia and 0.5350 in GRAPE. For AutoMorph it was 0.7481 and 0.5606, respectively. If we instead compute the Pearson correlation between λ and logit(worst quality), for DART correlations are 0.8570 and 0.5915 in the two datasets, and for AutoMorph 0.8941 and 0.6082 (all $P < 0.0001$). Thus the raw logits of QuickQualMEME's quality score are a better linear predictor of λ than the probability itself.

Figure 4 shows λ against the worst image quality in that eye. Cases of very high λ (>75%) all have very poor image quality (P(bad) > 0.8), and around P(bad) = 0.6 high measurement noise (>25%) appears to become more common. We can notice a visual difference between Caledonia and GRAPE. In GRAPE, there are cases of high λ even at good image quality and the correlation between λ and worst quality is lower. This is not unexpected, the long interval between images in GRAPE means that there could be due to changes retinal vasculature. However, λ is still clearly correlated with worst image quality. Thus, although there might be genuine changes in vasculature in GRAPE, cases of high λ are likely driven by poor image quality.

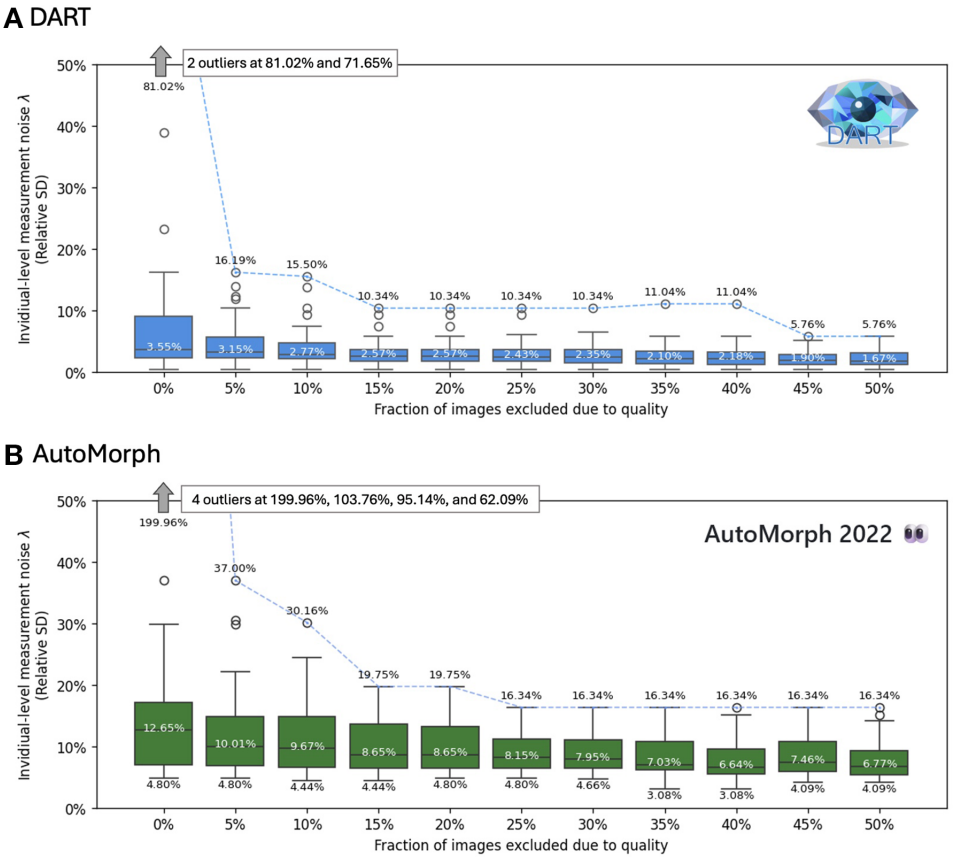


Figure 3. Individual-level measurement noise λ versus fraction of images excluded due to quality in the Caledonia dataset for DART (**A**) and AutoMorph (**B**). Lower is better. The top, middle, and bottom numbers indicate the highest, median, and lowest values, respectively. For DART, the lowest value was a constant 0.40% and omitted for space reasons. The *dashed blue line* joins the highest values. Both plots have the same scaling to allow for visual comparison. When no images are excluded, there are some outliers that are denoted on the plot.

Discussion

Both methods showed reasonable to good repeatability at the population level, even without any images being excluded. Interestingly, Pearson and Spearman correlations were comparable between Caledonia and GRAPE, despite the fact that GRAPE should provide a pessimistic lower-bound of performance for the reasons outlined in the Methods section. This is likely due to the low between-eyes SD of FD in Caledonia as subjects were relatively young and healthy. For DART FD, SD was five times higher in GRAPE and for AutoMorph 3.7 times. For a constant level of absolute measurement noise, smaller between-eyes SD will yield lower correlations.

DART showed higher repeatability than AutoMorph for all metrics, especially so on the Caledonia dataset. On the GRAPE dataset, both methods were more similar. This could be due to the long follow-up time, which means that differences in FD are a combination of genuine vascular changes and measurement noise, making differences in measurement noise between the two methods appear less pronounced.

At the individual level, repeatability in terms of λ was generally good, though there were some large outliers without quality exclusions. These outliers disappear even with modest levels of image quality exclusions. Repeatability improved generally as more images were excluded, but primarily affected large outliers.

Similar to the population-level metrics, DART had smaller λs than AutoMorph, both with and without quality exclusions. Interestingly, while robustness to image quality issues was a key motivation for DART's development, DART not only had smaller outliers at low levels of exclusions, but also clear advantage in best, median and worst case λ at any level of exclusions. Thus DART is also more repeatable in good quality images.

Based on the values of λ we observed in both datasets, both AutoMorph and DART might be applicable to individual-level risk prediction if we are targeting a population with large variation in FD, i.e. a more general population that is heterogeneous in age and systemic health, and especially if the expected effect on FD is large. The observed values of λ are generally small enough that we would rarely confuse high-, medium-, and low-FD individuals, especially when discarding images with very bad quality (i.e. QuickQual-MEME $P(\text{bad}) > 0.8$).

However, with median λ of 12.65% without exclusions and 6.64% with a high level of 40% of images excluded because of quality, AutoMorph would not be able to detect small changes (e.g., in a cohort with similar age and systemic health or when looking at longitudinal changes in an individual) and might not be useful for individual-level predictions if the effect on FD is small (e.g. for early-stage disease). DART on the other hand might be able to detect such small changes with a median λ of 3.55% without exclusions and

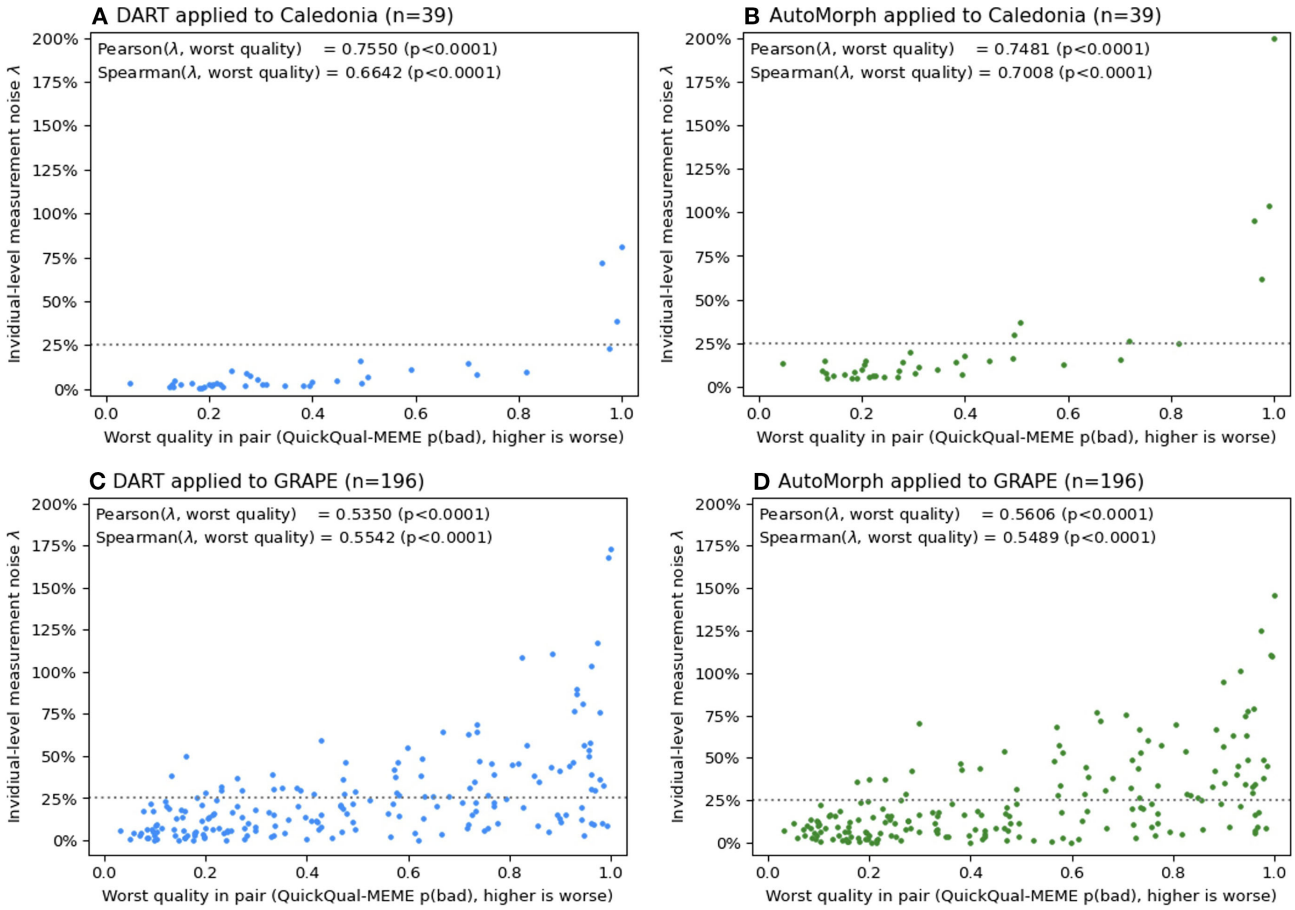


Figure 4. Individual-level measurement noise λ versus worst quality per eye for both methods in both datasets. Each point represents one eye. All plots have the same scaling to allow for visual comparison. The Pearson and Spearman correlation between λ and worst quality is reported in the top left corner of each plot. The *dashed horizontal line* indicates λ = 25% as a visual aid.

1.67% with exclusions, which would make it more useful for individual-level predictions and more appropriate for monitoring longitudinal changes.

Generally, these are encouraging results for the applicability of oculomics for individual-level predictions, especially considering the metrics on the GRAPE dataset provide a pessimistic lower bound because of its longitudinal nature. Although population and individual level is a common dichotomy in the literature, a more repeatable method is necessarily less noisy, and thus these results are also encouraging for population-level research. DART was more repeatable than AutoMorph even when excluding bad quality images, which highlights the value of designing robust methods for oculomics and retinal image analysis generally.

A key limitation of this work is the analyzed datasets. The GRAPE dataset is longitudinal and thus only provides a pessimistic lower bound of repeatability. On the other hand, the Caledonia dataset only contained healthy, relatively young adults and thus had low heterogeneity in FD. Additionally, there are endless alternative ways of analyzing the data at hand and further metrics that readers might be interested in.

Future work should examine repeatability of FD in additional, diverse cohorts. An ideal dataset for this would span a very wide age range, include diverse individuals with heterogeneous systemic health in different healthcare contexts, and have longitudinal data with multiple images per visit, so that measurement noise can be compared to longitudinal changes in the same individuals. Future work should also analyze the repeatability of additional tools such as VAMPIRE,[13] SIVA,[18] or IVAN,[19] as well as additional retinal traits like tortuosity and complexity index.[20,21]

Acknowledgments

The authors thank all participants in the studies used in this paper. We especially thank Kai Jin and Juan Ye as well as their colleagues for making the GRAPE dataset openly available to the research community.

JE was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) license to any author accepted manuscript version arising. M.O.B. gratefully acknowledges funding from: Fondation Leducq Transatlantic Network of Excellence (17 CVD 03); EPSRC grant no. EP/X025705/1; British Heart Foundation and The Alan Turing Institute Cardiovascular Data Science Award (C-10180357); Diabetes UK (20/0006221); Fight for Sight (5137/5138); the SCONe projects funded by Chief Scientist Office, Edinburgh & Lothians Health Foundation, Sight

Scotland, the Royal College of Surgeons of Edinburgh, the RS Macdonald Charitable Trust, and Fight For Sight.

Disclosure: **J. Engelmann,** None; **D. Moukaddem,** None; **L. Gago,** None; **N. Strang,** None; **M.O. Bernabeu,** None

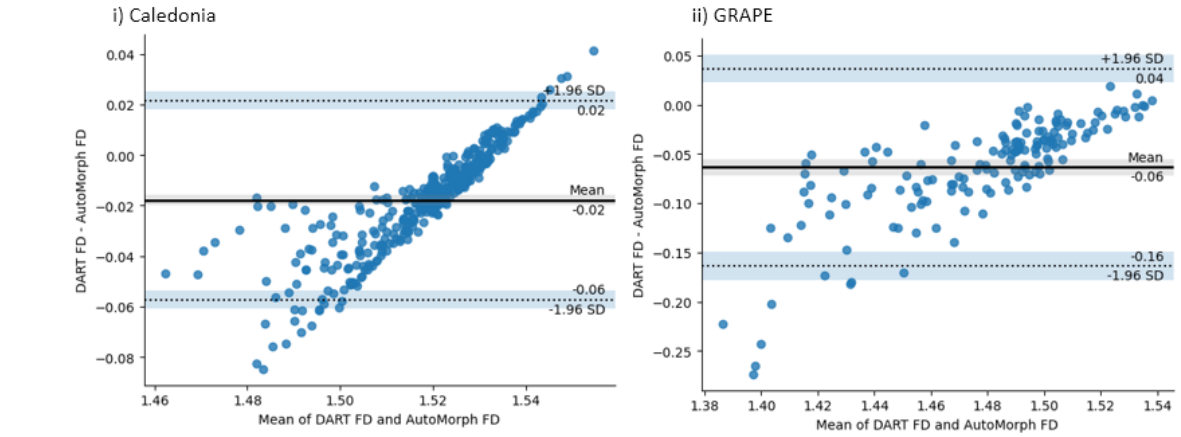
References

1. MacGillivray TJ, Trucco E, Cameron JR, Dhillon B, Houston JG, Van Beek EJR. Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *Br J Radiol*. 2014;87(1040):20130832.
2. Wagner SK, Fu DJ, Faes L, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol*. 2020;9(2):6–6.
3. Villaplana-Velasco A, Engelmann J, Rawlik K, et al. Decreased retinal vascular complexity is an early biomarker of MI supported by a shared genetic control. *medRxiv*. 2021.12. 16.21267446, 2021.
4. Mordi I, Trucco E. The eyes as a window to the heart: looking beyond the horizon. *Br J Ophthalmol*. 2022;106:1627, doi:10.1136/bjo-2022-322517.
5. Zekavat SM, Raghu VK, Trinder M, et al. Deep learning of the retina enables phenome-and genome-wide analyses of the microvasculature. *Circulation*. 2022;145(2):134–150.
6. McGrory S, Ballerini L, Doubal FN, et al. Retinal microvasculature and cerebral small vessel disease in the Lothian birth Cohort 1936 and mild stroke study. *Sci Rep*. 2019;9(1):6320.
7. Luben R, Wagner S, Struyven R, et al. Retinal fractal dimension in prevalent dementia: the AlzEye study. *Invest Ophthalmol Vis Sci*. 2022;63(7):4440-F0119-4440-F0119.
8. MacGillivray TJ, Cameron JR, Zhang Q, et al. Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS One*. 2015;10(5):e0127914.
9. Engelmann J, Storkey A, Linares MB. Exclusion of poor quality fundus images biases health research linking retinal traits and systemic health. *Invest Ophthalmol Vis Sci*. 2023;64:2922–2922.
10. Zhou Y, Wagner SK, Chia MA, et al. AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Transl Vis Sci Technol*. 2022;11(7):12.
11. Engelmann J, Villaplana-Velasco A, Storkey A, Bernabeu MO. Robust and efficient computation of retinal fractal dimension through deep approximation. In: *International Workshop on Ophthalmic Medical Image Analysis*. Berlin: Springer; 2022:84–93.
12. Huang X, Kong X, Shen Z, et al. GRAPE: a multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. *Sci Data*. 2023;10(1): 520.
13. Trucco E, Ballerini L, Relan D, et al. Novel VAMPIRE algorithms for quantitative analysis of the retinal vasculature. In *2013 ISSNIP Biosignals and Biorobotics Conference: biosignals and robotics for better and safer living (BRC)*. IEEE; 2013:1–4.
14. Mautuit T, Cunnac P, Cheung CY, et al. Concordance between SIVA, IVAN, and VAMPIRE Software Tools for semi-automated analysis of retinal vessel caliber. *Diagnostics*. 2022;12:1317.
15. McGrory S, Taylor AM, Pellegrini E, et al. Towards standardization of quantitative retinal vascular parameters: comparison of SIVA and VAMPIRE measurements in the Lothian Birth Cohort 1936. *Transl Vis Sci Technol*. 2018;7 (2):12.
16. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008;31:466–475.
17. Bishara AJ, Hittner JB. Confidence intervals for correlations when data are not normal. *Behav Res*. 2017;49:294–309.
18. Cheung CY, Tay WT, Mitchell P, et al. Quantitative and qualitative retinal microvascular characteristics and blood pressure. *J Hypertens*. 2011;29:1380–1391.
19. Klein R, Myers CE, Knudtson MD, et al. Relationship of blood pressure and other factors to serial retinal arteriolar diameter measurements over time: the beaver dam eye study. *Arch Ophthalmol*. 2012;130:1019–1027.
20. Alam M, Le D, Lim JI, Yao X. Vascular complexity analysis in OCT angiography of diabetic retinopathy. *Retina*. 2021;41:538–545.
21. Araya-Arriagada J, Garay S, Rojas C, et al. Multiscale entropy analysis of retinal signals reveals reduced complexity in a mouse model of Alzheimer's disease. *Sci Rep*. 2022;12(1):8900.

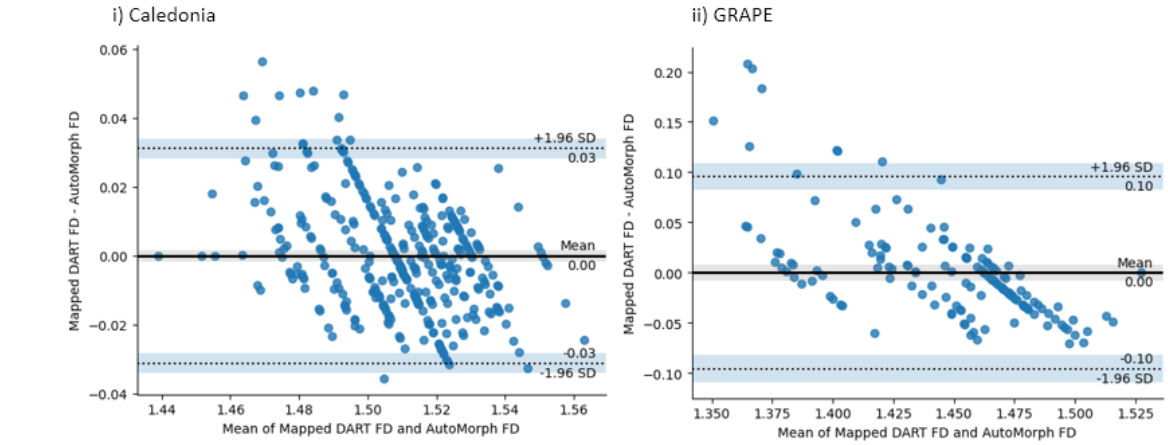
Supplementary

Supplementary Figure 1: Interchangeability of FD measurements by DART and AutoMorph. We use mean values per eye to reduce noise and exclude images with p(bad)>0.8. a) shows Bland-Altman plots using the original values for both Caledonia (i) and GRAPE (ii). b) shows Bland-Altman plots after DART FD values have been mapped to AutoMorph values using isotonic regression, to investigate whether it is possible to "translate" between the two tools. c) shows the isotonic regression lines and underlying data.

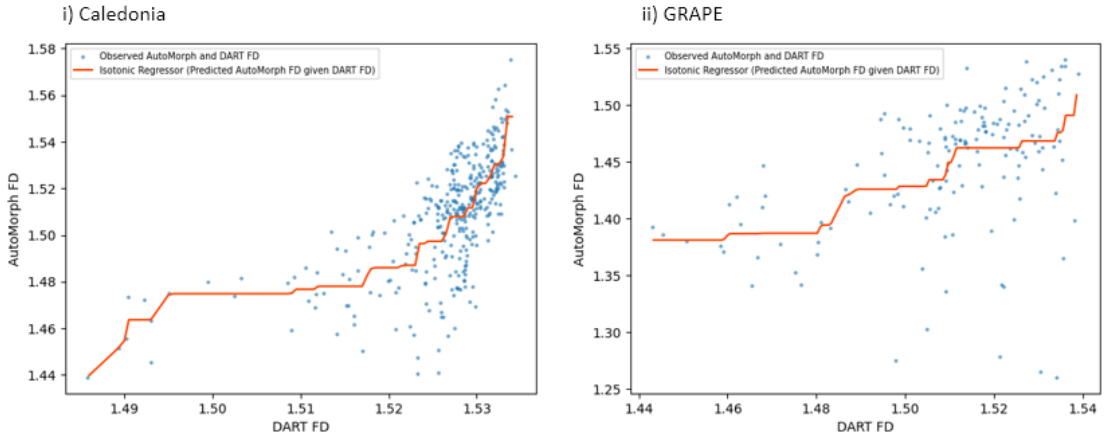
a) Bland-Altman plots using raw values



b) Bland-Altman plots using mapped DART FD values



c) Isotonic mapping



5.3 Conclusion

I think these results are quite encouraging regarding the potential of DART, indeed to a degree that surprised me when I analysed the data. DART is designed to be robust so I hoped and expected that it would be more repeatable than traditional approaches in the case of mixed image quality. However, I did not expect that it would be more repeatable even when poor quality images were excluded. Indeed, I think there was reason to think that it would be less repeatable for good quality images. One common criticism of DART is that it is a "black box" whereas the traditional approach of segmentation, refinement, box counting is more interpretable. I think this view has its limits, for instance previous work showed that different pipelines - each receiving the same high quality images as inputs - produced retinal traits that were the same in name but had poor to limited agreement (McGrory et al., 2018). To me this illustrates that while we might feel that each of the steps are easy to understand, there are differences in implementation as well as complex interactions between the different steps. Something comprised of multiple things that are individually simple can itself be highly complex, and for example a small change in vessel segmentation could propagate through the subsequent steps and end up having a large effect on the final output in a way that is hard to anticipate from studying each step in isolation.

From a theoretical perspective, the results are not entirely surprising. DART learned to "map" from images to fractal dimension and during that process was encouraged to ignore rather substantial variations in image quality and map different versions of the same image to the same number. This process likely also encouraged ignoring small variations in image quality and mapping slightly different versions of the same image (i.e. multiple pictures of the same eye during the same visit) to the same number. Still, I did not expect that DART would be more repeatable for high quality images, too. When analysing the data, I was very careful and mindful of avoiding anything that might bias the results in favour of DART, since it is a method I developed myself. I think the results are reliable and accurate, but I hope that in the future other researchers independently compare DART with other approaches to validate my findings.

In the future, this analysis should be repeated on larger, more diverse repeated dataset. The repeated dataset from Glasgow Caledonian University was of young, healthy adults who had relatively high fractal dimension and generally very high image quality. The GRAPE dataset from Zhejiang University (Huang et al., 2023) was more diverse in terms of age, but only included glaucoma patients and due to its longitudinal

nature only provided a lower bound for repeatability. Additionally, future work should investigate whether improved repeatability and robustness indeed translate to stronger statistical associations and increased predictive power. It would also be interesting to examine repeatability and robustness of other retinal image analysis tools. Finally, now that we have some estimates for repeatability, it would be interesting to investigate whether there is any meaningful level of diurnal variation of fractal dimension, which future studies might then adjust for.

Chapter 6

Machine learning for automated analysis of the choroid in optical coherence tomography images

6.1 Introduction

The choroid is a highly vascular tissue situated between the retina and the sclera, the outer white part of the eye, that supplies the outer retina with blood. This location makes it hard to image and thus traditionally it has not been studied as much in relation to systemic health as the retinal vasculature that can be seen in colour fundus images. With the advent of and improvements in optical coherence tomography, including enhanced depth imaging (Spaide et al., 2008), the choroid can now be imaged easily and is increasingly studied in relation to cardiovascular (Yeung et al., 2020), neurodegenerative (Kundu et al., 2023; Robbins et al., 2021), and renal (Balmforth et al., 2016; Burke et al., 2023b; Shin et al., 2019) conditions. The choroid is of course also of interest in ophthalmology, e.g. in relation to myopia (Read et al., 2019) or central serous chorioretinopathy (Semeraro et al., 2019).

Quantitative analysis of the choroid through retinal traits such as choroidal thickness or vascular index (Iovino et al., 2020) requires segmentation of the choroid region and the vasculature within it. Manual segmentation requires experienced graders, is somewhat time-consuming for the region and almost prohibitively so for the vasculature, and introduces subjectivity. Semi-automatic methods (Burke and King, 2021; Eghtedar et al., 2022) only partially alleviate these issues. For region segmentation, fully-automatic methods have been proposed but require a good understanding of image processing and proprietary software to use (Mazzaferri et al., 2017), or do not currently appear to openly available (Kugelman et al., 2019; Xuan et al., 2023). For

vessel segmentation, there are thresholding-based methods (Agrawal et al., 2020). However, even different implementations of the same thresholding method might have poor agreement (Wei et al., 2018). Furthermore, due to variations in the optical coherence tomography images, due to image quality, scan settings, or device manufacturer, thresholding methods might require some adjustment of their parameters, which reintroduces subjectivity and the need for technical knowledge. Some deep learning-based methods for choroidal vessel segmentation have been proposed (Liu et al., 2019; Muller et al., 2022), but again do not currently appear to be openly available. Easy-to-use, openly available tools for automatic analysis of the choroid could help accelerate this nascent field of study.

This work was undertaken with a dear colleague of mine and fellow PhD student, Jamie Burke, whose PhD is focused on choroidal analysis. Despite both being at the same university, we only properly met when attending at ARVO annual meeting, a leading ophthalmology conference. At that point, Jamie had already developed a semi-automatic method for choroidal region segmentation called Gaussian Process Edge Tracing (Burke and King, 2021), or GPET for short, which required about half a minute of time and some manual inputs for each optical coherence tomography B-scan. I suggested that a deep learning model for segmentation could likely make this task fully-automatic and after returning to Edinburgh we started working on this project. Our collaboration already resulted in two peer-reviewed publications and substantial knowledge transfer between us.

Of course, this thesis is titled “machine learning for retinal image analysis” yet strictly speaking the choroid is distinct from the retina. Posterior segment optical coherence tomography is commonly referred to as a type of retinal imaging, while “ocular posterior segment imaging” is not a particularly common term. In terms of both methods and applications, this work does fit very well with this thesis.

6.2 Papers

6.2.1 Contributions

As mentioned in section 1.2 of this thesis as well as the introduction to this chapter, this is work I have jointly undertaken with my dear colleague Jamie Burke, who is likewise currently a PhD student. We are “joint first authors” on both manuscripts, while alternating the ordering of our names. Jamie’s PhD is focused on choroidal analysis and he has

deep expertise regarding both the choroid and optical coherence tomography imaging. Prior to working with me, he had already developed a semi-automatic method for choroidal region segmentation and code for calculation of choroidal area and thickness. He also curated relevant datasets. The work in this chapter builds on his work. I primarily contributed expertise regarding deep learning, writing the code for the deep learning part, and trained, validated and selected the models. However, while working together, we collaborated closely on all parts of the work and exchanged knowledge, so that now Jamie is very proficient in training deep learning models himself, while I learned a lot more about the choroid and optical coherence tomography. In addition to the deep learning part, I also had substantial contributions regarding the study design and key decisions regarding the data curation and evaluation, analysis of the results, as well as writing of the manuscript. Jamie and I led both pieces of work, with our collaborators primarily providing guidance, data, clinical expertise, and assisting in the manual evaluation of our models and baselines. Thus, in my opinion, both pieces of work are worthy of inclusion in the present thesis, just as they will be worthy of inclusion in Jamie's thesis.

6.2.2 First paper

tvst

Artificial Intelligence

An Open-Source Deep Learning Algorithm for Efficient and Fully Automatic Analysis of the Choroid in Optical Coherence Tomography

Jamie Burke[1,*], Justin Engelmann[2,3,*], Charlene Hamid[4], Megan Reid-Schachter[4], Tom Pearson[5], Dan Pugh[6], Neeraj Dhaun[6], Amos Storkey[7], Stuart King[1], Tom J. MacGillivray[4,8], Miguel O. Bernabeu[3,9], and Ian J. C. MacCormick[7,10]

[1] School of Mathematics, University of Edinburgh, Edinburgh, UK
[2] School of Informatics, University of Edinburgh, Edinburgh, UK
[3] Centre for Medical Informatics, University of Edinburgh, Edinburgh, UK
[4] Clinical Research Facility and Imaging, University of Edinburgh, Edinburgh, UK
[5] University Hospital Wales, NHS Wales, Cardiff, Wales, UK
[6] British Heart Foundation Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK
[7] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, UK
[8] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[9] The Bayes Centre, University of Edinburgh, Edinburgh, UK
[10] Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK

Correspondence: Jamie Burke, School of Mathematics, University of Edinburgh, College of Science and Engineering, Edinburgh, UK. e-mail: jamie.burke@ed.ac.uk

Received: August 22, 2023
Accepted: October 24, 2023
Published: November 21, 2023

Keywords: choroid; optical coherence tomography; deep learning; segmentation

Citation: Burke J, Engelmann J, Hamid C, Reid-Schachter M, Pearson T, Pugh D, Dhaun N, Storkey A, King S, MacGillivray TJ, Bernabeu MO, MacCormick IJC. An open-source deep learning algorithm for efficient and fully automatic analysis of the choroid in optical coherence tomography. Transl Vis Sci Technol. 2023;12(11):27, https://doi.org/10.1167/tvst.12.11.27

Purpose: To develop an open-source, fully automatic deep learning algorithm, DeepGPET, for choroid region segmentation in optical coherence tomography (OCT) data.

Methods: We used a dataset of 715 OCT B-scans (82 subjects, 115 eyes) from three clinical studies related to systemic disease. Ground-truth segmentations were generated using a clinically validated, semiautomatic choroid segmentation method, Gaussian Process Edge Tracing (GPET). We finetuned a U-Net with the MobileNetV3 backbone pretrained on ImageNet. Standard segmentation agreement metrics, as well as derived measures of choroidal thickness and area, were used to evaluate DeepGPET, alongside qualitative evaluation from a clinical ophthalmologist.

Results: DeepGPET achieved excellent agreement with GPET on data from three clinical studies (AUC = 0.9994, Dice = 0.9664; Pearson correlation = 0.8908 for choroidal thickness and 0.9082 for choroidal area), while reducing the mean processing time per image on a standard laptop CPU from 34.49 $\pm$ 15.09 seconds using GPET to 1.25 $\pm$ 0.10 seconds using DeepGPET. Both methods performed similarly according to a clinical ophthalmologist who qualitatively judged a subset of segmentations by GPET and DeepGPET, based on smoothness and accuracy of segmentations.

Conclusions: DeepGPET, a fully automatic, open-source algorithm for choroidal segmentation, will enable researchers to efficiently extract choroidal measurements, even for large datasets. As no manual interventions are required, DeepGPET is less subjective than semiautomatic methods and could be deployed in clinical practice without requiring a trained operator.

Translational Relevance: DeepGPET addresses the lack of open-source, fully automatic, and clinically relevant choroid segmentation algorithms, and its subsequent public release will facilitate future choroidal research in both ophthalmology and wider systemic health.

Introduction

The retinal choroid is a complex, extensively interconnected vessel network positioned between the retina and the sclera. The choroid holds the majority of the vasculature in the eye and plays a pivotal role in nourishing the retina. Optical coherence tomography (OCT) is an ocular imaging modality that uses low-coherence light to construct a three-dimensional map of chorioretinal structures at the back of the eye. Standard OCT imaging does not visualize the deeper choroidal tissue well, as the tissue sits beneath the hyperreflective retinal pigment epithelium layer of the retina. Enhanced depth imaging OCT (EDI-OCT) overcomes this problem and offers improved visualization of the choroid, thus providing a unique window into the microvascular network that not only resides closest to the brain embryologically but also carries the highest volumetric flow per unit tissue weight compared to any other organ in the body.

Since the advent of OCT, interest in the role played by the choroid in systemic health has been growing,[1] as non-invasive imaging of the choroidal microvasculature may provide a novel location to detect systemic, microvascular changes early. Indeed, changes in choroidal blood flow, thickness, and other markers have been shown to correspond with patient health such as choroidal thickness in chronic kidney disease[2] and choroidal area and vascularity in Alzheimer's dementia.[3]

Quantification of the choroid in EDI-OCT imaging requires segmentation of the choroidal space. However, this is a more difficult problem than retinal layer segmentation due to poor signal penetration from the device—and thus lower signal-to-noise ratio—and shadows cast by superficial retinal vessels and choroidal stroma tissue. This results in poor intra- and interrater agreement even with manual segmentation by experienced clinicians, and manual segmentation is too labor intensive and subjective to be practical for analyzing large-scale datasets.

Semiautomated algorithms improve on this slightly but are typically multistage procedures, requiring traditional image processing techniques to prepare the images for downstream segmentation.[4] Methods based on graph theory such as Dijkstra's algorithm[5,6] or graph cut,[7] as well as on statistical techniques including level sets,[8,9] contour evolution,[10] and Gaussian mixture models,[11] have been proposed previously. Concurrently, deep learning (DL)-based approaches have emerged. Chen et al.[12] used a DL model for choroid layer segmentation but with traditional contour tracing as a postprocessing step. Other DL-based approaches, too, combine traditional image processing techniques as pre- or postprocessing steps,[13–15] whereas others are fully DL based,[16,17] the latter of which is in a similar vein to the proposed method. More recently, DL has been used to distill existing semiautomatic traditional image processing pipelines into a fully automatic method.[18]

Gaussian Process Edge Tracing (GPET), based on Bayesian machine learning,[19] is a particularly promising method for choroid layer segmentation that has been clinically and quantitatively validated.[20] Gaussian process regression is used to model the upper and lower boundaries of the choroid from OCT scans. For each boundary, a recursive Bayesian scheme is employed to iteratively detect boundary pixels based on the image gradient and the distribution of candidate boundaries by the Gaussian process regressor. However, GPET is semiautomatic and thus requires time-consuming manual interventions by specifically trained personnel, which introduces subjectivity and limits the potential for analyzing larger datasets or deploying GPET into clinical practice.

There are currently no accessible, open-source algorithms for fully automatic choroidal segmentation. All available algorithms fall into one of three categories: First, semiautomatic methods[21,22] that require human supervision and training and introduce subjectivity. Second, fully automatic DL-based methods that are not openly accessible, either providing only the code but not the trained model necessary to use the method[23] or not providing any access at the time of writing.[24] Third, fully automatic algorithms comprising many steps that require a good understanding of image processing techniques and a license for proprietary software (MATLAB; MathWorks, Natick, MA).[25]

We aimed to develop and release an open-source, raw image-to-measurement, fully automatic method for choroid region segmentation that can be easily used without special training and does not require licenses for proprietary software (Fig. 1). Importantly, we intend not only to make our method available to the research community but also to do so in a frictionless way that allows other researchers to download and use our method without seeking our approval. We distill GPET into a DL algorithm, DeepGPET, which can process images without supervision in a fraction of the time, permitting analysis of large-scale datasets and potential deployment into clinical care and research practice without prior training in image processing. The code and model weights for DeepGPET are available at https://github.com/jaburke166/deepgpet.

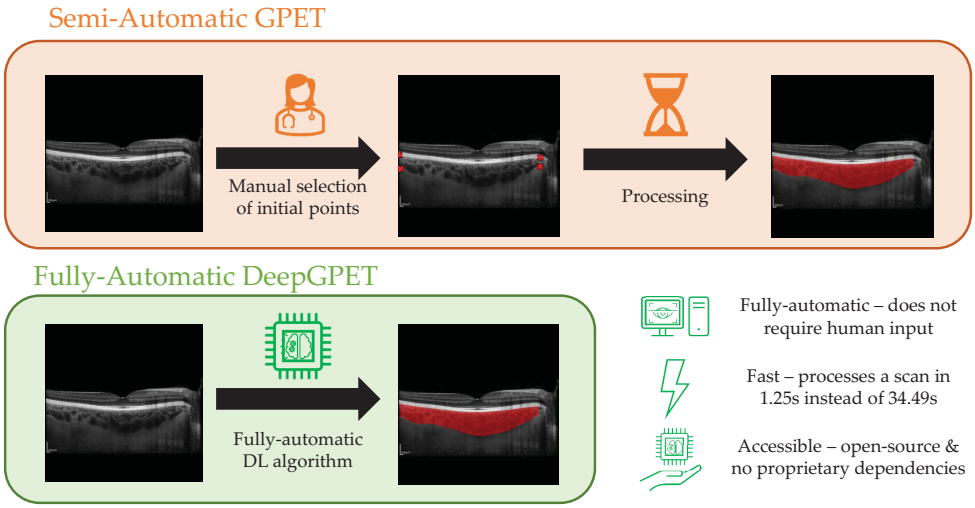


Figure 1. Comparison between the semi-automatic GPET[19,20] (*top*) and fully automatic DeepGPET (*bottom*).

Methods

Study Population

We used 715 OCT B-scans belonging to 82 subjects from three studies: (1) OCTANE,[26] a study looking at renal function and impairment in chronic kidney disease patients; (2) i-Test, a study recruiting pregnant women of any gestation or those who have delivered a baby within 6 months, including controls and individuals at high risk of complications; and (3) normative data from 30 healthy volunteers as a control group.[27] All studies conformed to the tenets of the Declaration of Helsinki and received relevant ethical approval and informed consent from all subjects. Table 1 provides an overview of basic population characteristics and number of subjects and images for these studies. Supplementary Figure S1 presents boxplot distributions of choroidal thickness and area for the three datasets used to build DeepGPET, and Table 1 presents tabular mean and standard deviation values.

Two Heidelberg Engineering (Heidelberg, Germany) spectral-domain OCT SPECTRALIS devices were used for image acquisition: the Standard Module (OCT1 system) and FLEX Module (OCT2 system). The FLEX is a portable version that enables imaging of patients in a ward environment. Both machines image a 30° region (8.7 mm), generating a macular, cross-sectional OCT B-scan at a resolution of 768 × 768 pixels. Notably, 14% of the OCT B-scans were scanned without EDI mode activated (non-EDI) and thus presented more challenging images with lower signal-to-noise ratios in the choroidal part of the OCT. Horizontal line and vertical scans were centered at the fovea with active eye tracking using an Automatic Real Time (ART) value of 100. Posterior pole macular scans covered a 30° × 25° region using EDI mode.

We split the data into an approximately 85:8:7 split among training (603 B-scans, 66 subjects), validation

Table 1. Overview of Population Characteristics

	OCTANE	i-Test	Control	Total
Subjects, *n*	47	5	30	82
Male/female, *n*	24/23	0/5	20/10	44/38
Right/left eyes, *n*	47/0	5/5	29/29	81/34
Age (y), mean (SD)	48.8 (12.9)	34.4 (3.4)	49.1 (7.0)	48.0 (11.2)
Machine	Standard	FLEX	Standard	Both
Horizontal/vertical scans, *n*	166/0	16/16	57/54	239/70
Volume scans, *n*	174	186	46	406
Total B-scans, *n*	340	218	157	715

(58 B-scans, 9 subjects), and test sets (54 B-scans, 7 subjects). When splitting the data, we did so at the patient level; that is, each subject's OCT images were present in only one set and they were selected so that each set had proportionally equal amounts of scan types (EDI/non-EDI) to best represent image quality. See Supplementary Table S1 for an overview of basic population and imaging characteristics for each set.

DeepGPET

As the ground truths are based on GPET, DeepGPET can be can be seen as a more efficient, fully automatic and distilled version of GPET. Our approach was to fine-tune a U-Net[28] with a MobileNetV3[29] backbone pretrained on ImageNet for 60 epochs with batch size 16 using AdamW[30] (learning rate [LR] = 10^{-3}; $\beta_1 = 0.9$; $\beta_2 = 0.999$; weight decay = 10^{-2}). After epoch 30, we maintained an exponential moving average of model weights which we then used as our final model. We used the following data augmentations: brightness and contrast changes, horizontal flipping, and simulated OCT speckle noise by applying Gaussian noise followed by multiplicative noise (all $P = 0.5$), and Gaussian blur and random affine transforms (both $P = 0.25$). To reduce memory load, we cropped the black space above and below the OCT B-scan and processed the images at a resolution of 544 × 768 pixels. Images were standardized by subtracting 0.1 and dividing by 0.2, and no further preprocessing was done. We used Python 3.11, PyTorch 2.0, Segmentation Models PyTorch,[31] and the timm library.[32]

Statistical Analysis

We used the Dice coefficient and area under the receiver operating characteristic curve (AUC) for evaluating agreement in segmentations, as well as the Pearson correlation coefficient (r) and mean absolute error (MAE) for segmentation-derived choroid thickness and area. The calculation of thickness and area from the segmentation is described in more detail in Burke et al.[20] Briefly, for thickness, the average of three measures was used, taken at the fovea and 2000 µm from it in either direction by drawing a perpendicular line from the upper boundary to the lower boundary to account for choroidal curvature. For area, pixels were counted in a region of interest with radius 3000 µm around the fovea, corresponding to the commonly used Early Treatment Diabetic Retinopathy Study (ETDRS) macular area of 6000 × 6000 µm.[33]

We compared the agreement of DeepGPET with GPET segmentations against the repeatability of GPET itself. The creator of GPET (J.B.) made both the original and repeated segmentations with GPET. Because both segmentations were done by the same person there was no interrater subjectivity at play. Thus, the intrarater agreement measured here is a best-case scenario and forms an upper bound for agreement with the original segmentations and any other semiautomatic method requiring manual input, which can necessarily be subject to human variability, unlike DeepGPET.

In addition to quantitative evaluations, we also compared segmentations by GPET and DeepGPET for 20 test-set OCT images qualitatively by having them rated by an experienced clinical ophthalmologist (I.J.C.M.). We selected seven examples with the highest disagreement in thickness and area, seven examples with disagreement closest to the median, and six examples with the lowest disagreement. Thus, these 20 examples cover cases where both methods are very different, cases of typical disagreement, and cases where both methods are very similar. In each instance, the ophthalmologist (I.J.C.M.) was shown the segmentations of both methods overlaid on the OCT (blinded to which method produced which segmentation) and was also provided with the raw, full-resolution OCT. He was then asked to rate each one along three dimensions: quality of the upper boundary, quality of the lower boundary and overall smoothness using an ordinal scale of "very bad", "bad", "okay", "good", or "very good."

Results

Quantitative

Table 2 shows the results for DeepGPET and a repeat GPET compared to the initial GPET segmentation as "ground truth."

Agreement in Segmentation

Both methods show excellent agreement with the original segmentations. The agreement of DeepGPET is comparable to the repeatability of GPET itself, with the DeepGPET AUC being slightly higher (0.9994 vs. 0.9812) and the Dice coefficient slightly lower (0.9664 vs. 0.9672). DeepGPET performing better in terms of AUC but worse in terms of Dice suggests that, for pixels where it disagreed with GPET after thresholding, the confidence is lower than for ones where it agreed with GPET. This in turn suggests that DeepGPET is well calibrated based on the raw predictions made for each pixel.

Table 2. Metrics for DeepGPET and Repeated GPET Using the Initial GPET Annotation as Ground Truth

				Thickness		Area	
Method	AUC	Dice	Time (s/image), Mean ± SD	Pearson's *r*	MAE (μm)	Pearson's *r*	MAE (mm^2)
DeepGPET	0.9994	0.9664	1.25 ± 0.10	0.8908	13.3086	0.9082	0.0699
Repeat GPET	0.9812	0.9672	34.49 ± 15.09	0.9527	10.4074	0.9726	0.0486

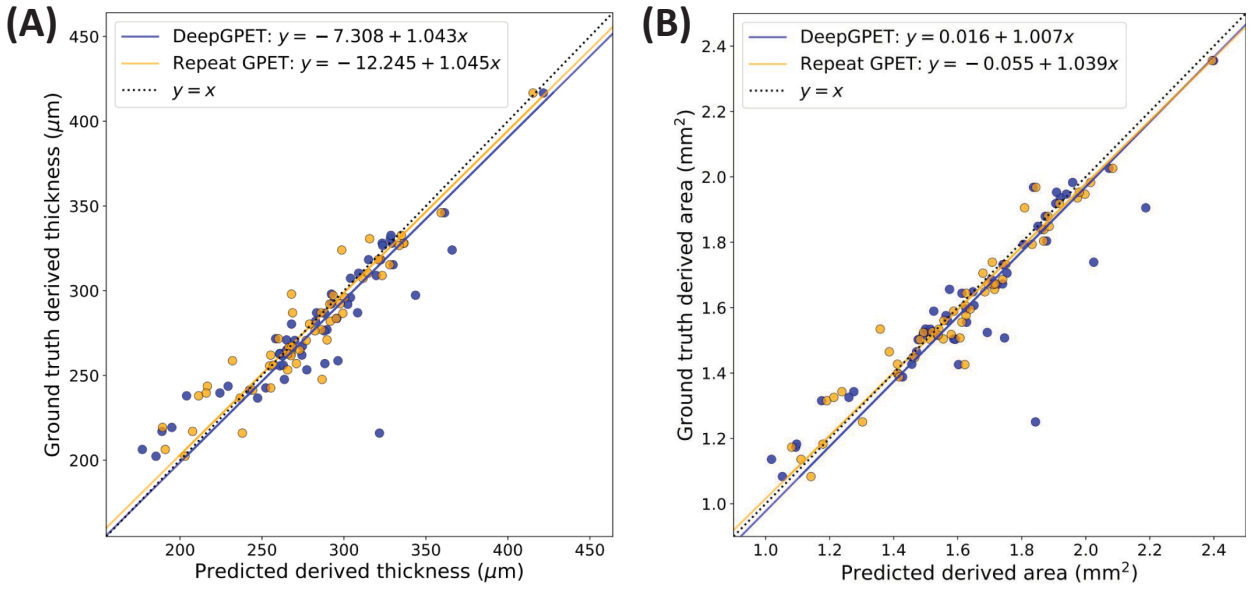


Figure 2. Correlation plots comparing derived measures of mean choroid thickness (**A**) and choroid area (**B**) using DeepGPET and the resegmentations using GPET.

Processing Speed and Manual Interventions

Both methods were compared on the same standard laptop CPU. DeepGPET processed the images in only 3.6% of the time that GPET required. DeepGPET was fully automatic, and it successfully segmented all images, whereas GPET required 1.27 manual interventions on average, including selecting initial pixels and manual adjustment of GPET parameters when the initial segmentation failed.

This faster processing results in massive time savings. A standard OCT volume scan consists of 61 B-scans. With GPET, processing such a volume for a single eye takes about 35 minutes, during which a person has to select initial pixels to guide tracing (for all images) and adjust parameters if GPET initially failed (for about 25% of images). In contrast, DeepGPET could do the same processing in about 76 seconds on the same hardware, during which no manual input is needed. DeepGPET could even be GPU accelerated to cut the processing time by another order of magnitude.

The lack of manual interventions required by DeepGPET means that no subjectivity is introduced, unlike GPET, particularly when used by different people. Additionally, DeepGPET does not require specifically trained analysts and could be used fully automatically in clinical practice.

Agreement in Choroid Area and Thickness

GPET showed very high repeatability for thickness (Pearson's $r = 0.9527$; MAE = 10.4074 μm) and area (Pearson's $r = 0.9726$; MAE = 0.0486 mm^2). DeepGPET achieved slightly lower, yet also very high agreement for both thickness (Pearson's $r = 0.8908$; MAE = 13.3086 μm) and area (Pearson's $r = 0.9082$; MAE = 0.0699 mm^2). Figure 2 shows correlation plots for thickness and area. The agreement between DeepGPET and GPET did not quite reach the repeatability of GPET itself when used by the same experienced analyst, but it was quite comparable and high in absolute terms. Especially noteworthy is that the MAE for thickness and area was only 21% lower for thickness and 30% lower for area for repeated GPET than for DeepGPET. Thus, DeepGPET comes quite close to optimal performance (i.e., best-case repeatability where the same experienced analyst did both sets of annotation).

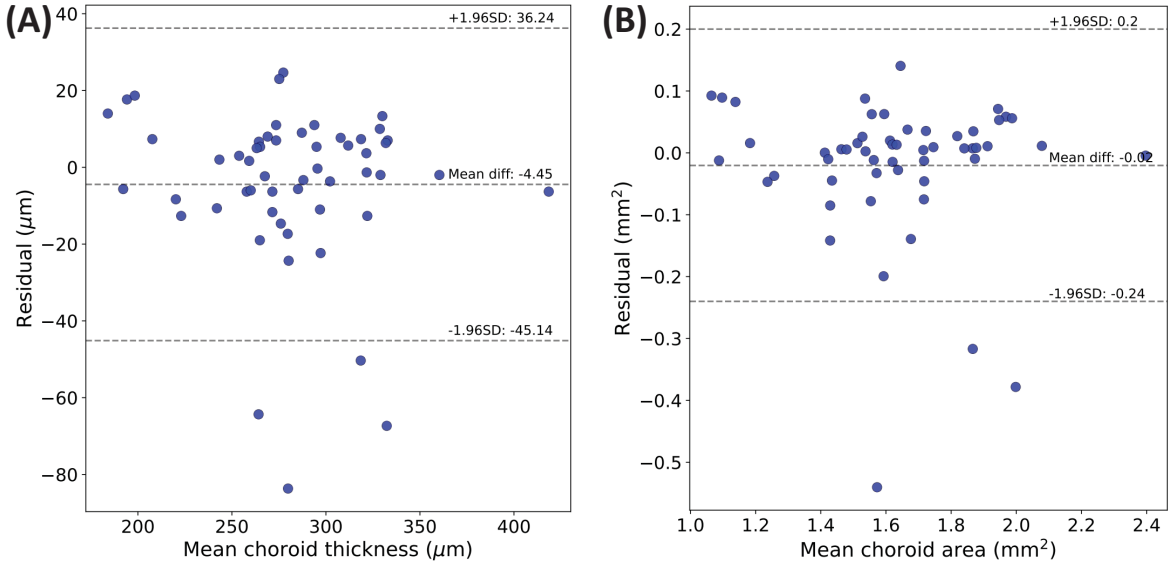


Figure 3. Bland–Altman plots comparing the agreement between DeepGPET and GPET using mean choroid thickness (**A**) and choroid area (**B**).

Table 3. Qualitative Ratings of 20 Test Set Segmentations Along Three Key Dimensions

Method	Upper Boundary	Lower Boundary	Smoothness
DeepGPET	Very good: 20	Very good: 4; good: 10; okay: 4; Bad: 2	Very good: 5; good: 12; okay: 2; bad: 1
GPET	Very good: 20	Very good: 6; good: 6; okay: 8; Bad: 0	Very good: 6; good: 13; okay: 1; bad: 0

The rater was blinded to the identity of the methods, and their order was randomized for every example.

Furthermore, the regression fits in both derived measures for DeepGPET are closer to the identity line than for the repeated GPET measurements. For choroid thickness (CT), the linear fit estimated a slope value of 1.043 (95% confidence interval [CI], 0.895–1.192) and intercept of −7.308 µm (95% CI, −48.967 to 34.350). For choroid area (CA), the linear fit estimated a slope value of 1.01 (95% CI, 0.878–1.137) and an intercept of 0.016 mm^2 (95% CI, −0.195 to 0.226). All CIs contain 1 and 0 for the slope and intercepts, respectively, suggesting no systematic bias or proportional difference between GPET and DeepGPET.[34,35]

Figure 3 shows the residuals between DeepGPET and the ground-truth labels from the held-out test set using Bland–Altman plots.[36] Rahman et al.[37] found that intrarater agreement and interrater agreement of subfoveal choroidal thickness measurements were 23 µm and 32 µm, respectively. For CT, only 9.3% were greater than 23 µm in absolute value (5/54), with four of these representing major sources of disagreement. Similarly, for CA, the majority of residuals were centered on 0 (mean residual of −0.02 mm^2), with only 5.5% of residuals (3/54) lying outside the limits of agreement.

Qualitative

Table 3 shows the results of the adjudication between GPET and DeepGPET. The upper boundary was rated as "very good" for both methods in all 20 cases. However, for the lower boundary, DeepGPET was rated as "bad" in two cases for the lower boundary and one case for smoothness. Otherwise, both methods performed very similarly.

Figure 4 shows some examples. Figure 4A shows that DeepGPET segmented more of the temporal region than did GPET, thus providing a full-width segmentation that was preferred by the rater. Additionally, both approaches are able to segment a smooth boundary, even in regions with stroma fluid obscuring the lower boundary (red arrow). In Figure 4B, the lower boundary for this choroid is very faint and is actually below the majority of the vessels sitting most posterior (red arrow). DeepGPET produced a smooth and concave boundary preferred by the rater, whereas GPET fell victim to hugging the posterior-most vessels in the subfoveal region. In Figure 4C, DeepGPET rejected the true boundary in the low-contrast region (red arrow) and opted for a more well-defined one,

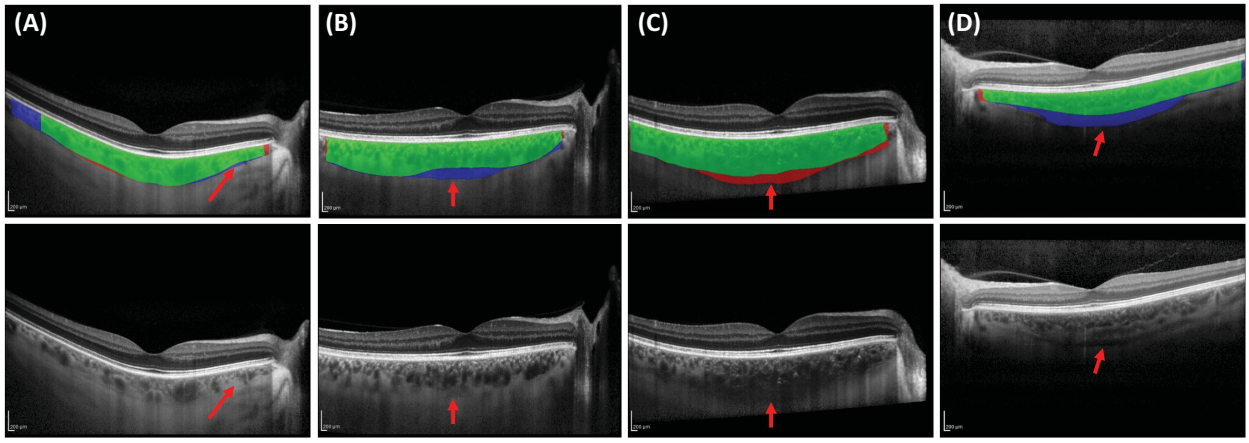


Figure 4. Four examples from the adjudication. The rater preferred DeepGPET for panels (**A**) and (**B**) and GPET for panels (**C**) and (**D**). (*Top row*) *Green* indicates segmentation by both GPET and DeepGPET; *red*, by GPET only; and *blue*, by DeepGPET only. (*Bottom row*) The *arrows* indicate important choroidal features that can make segmentation challenging. (**A**) No large vessels are in the nasal region to guide segmentation. (**B**) Lower boundary is very faint and below the posterior most vessels. (**C**) Lower boundary is noisy and faint. (**D**) Large suprachoroidal space is visible.

whereas GPET segmented the more uncertain path. Because GPET permits human intervention, there is more opportunity to finetune its parameters to fit what the analyst believes is the true boundary. Here, the rater preferred GPET, whereas the under-confidence of DeepGPET led to undersegmentation and to a bad rating. In Figure 4D, the lower boundary is difficult to delineate due to a thick suprachoroidal space (red arrow) and thus a lack of lower boundary definition. Here, the rater gave a bad rating to DeepGPET and preferred GPET, but remarked that GPET actually under-segmented the choroid by intersecting through posterior vessels. The choroids in Figures 4B to 4D are the choroids with the largest CT and CA disagreement between DeepGPET and GPET as observed in Figure 3.

Discussion

We developed DeepGPET, a fully automatic and efficient method for choroid layer segmentation, by distilling GPET, a clinically validated semiautomatic method. DeepGPET achieved excellent agreement with GPET on held-out data in terms of segmentation and derived choroidal measurements, approaching the repeatability of GPET itself and well within the threshold expected to exceed interrater agreement as observed in previous work.[37] We also found no significant association between segmentation performance (via Dice score) and choroidal thickness or area and the Heidelberg signal-to-noise quality index in the held-out test set (Supplementary Table S3 and Supplementary Fig. S2). Most importantly, DeepGPET does not require specialist training and can process images fully automatically in a fraction of the time, enabling analysis of large-scale datasets and potential deployment in clinical practice.

Although the observed agreement was very high, it was not perfect. However, even higher agreement with GPET would not necessarily produce a better method, as GPET itself is not perfect and even conceptually there is debate around the exact location of the choroid–scleral interface (CSI), the lower choroid boundary in an OCT B-scan. CSI is commonly defined (e.g., by the original authors behind EDI-OCT[38]) as the smooth inner boundary between the choroid and sclera, or just below the most posterior vessels but excluding the suprachoroidal space. However, even that definition is still debated and can be difficult to discern in practice. Not all choroids are smooth, and there are edge cases such as vessels passing from the sclera into the choroid or stroma fluid obscurations that make the boundary even more ambiguous. These features, coupled with low signal-to-noise ratio and vessel shadowing from superficial retinal vessels, all contribute to the difficult challenge of choroid layer segmentation.

For quantitative analysis of choroidal phenotypes, the specific definition of the CSI is secondary to applying the same, consistent definition across and within patients. Here, fully automatic methods such as DeepGPET provide great benefit by removing the subjectivity present in semiautomatic methods. Where semiautomatic methods require manual input, two

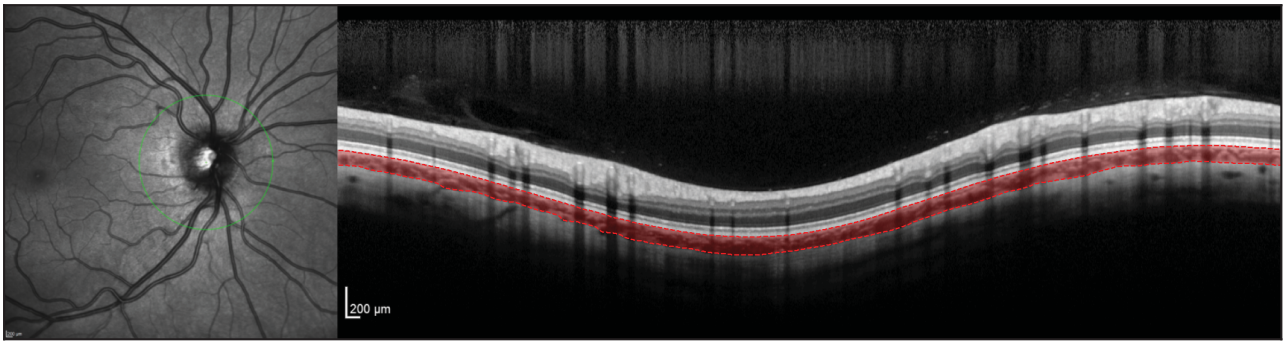


Figure 5. An example peripapillary scan from the Heidelberg Standard Module, automatically segmented by DeepGPET without manual intervention.

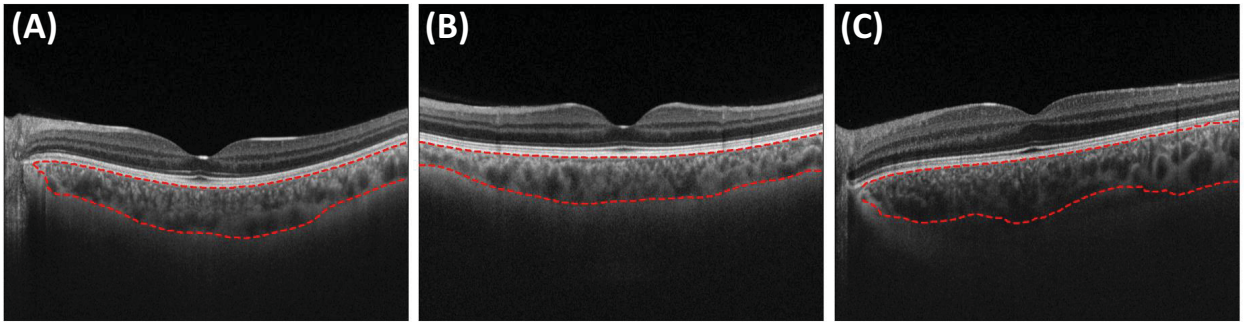


Figure 6. Three OCT B-scan images from a Topcon imaging device, of which two were successful (**A**, **B**) and one was not (**C**).

analysts with different understandings of the CSI could produce vastly different segmentations. With DeepGPET, the same image is always segmented in the same way, removing subjectivity.

Initial experiments with other types of OCT imaging have positively indicated the ability of DeepGPET to generalize to different visualizations of the choroid. Figure 5 shows a peripapillary scan extracted from the Heidelberg Standard Module, centered on the optic head, with the choroid automatically segmented. Figure 6 shows choroid segmentations using DeepGPET for three OCT B-scans from a Topcon device (DRI OCT Triton Plus; Topcon, Tokyo, Japan)—two cases where DeepGPET worked well and one case where it did not. This result shows some promise in the usability of DeepGPET for scans different from the Heidelberg macular line scans on which it was trained. We hope in future iterations to extend the training data with scans from different imaging devices and scan locations. We recommend those using DeepGPET on non-Heidelberg images to review the segmentations afterward as a sanity check.

In the present work, we used data from three studies and two OCT devices and included both EDI and non-EDI scans. However, we only used data from subjects that were either healthy or had systemic but not eye disease, for which DeepGPET might not be robust. In future work, we plan to externally validate DeepGPET and include cases of ocular pathologies. A further limitation is that, although GPET has been clinically validated, not all segmentations used for training DeepGPET were entirely perfect. Thus, revisiting some of the existing segmentations and manually improving them to a "gold standard" for purposes of training the model could improve DeepGPET. For example, GPET does not always segment the entire width of the choroid. Interestingly, DeepGPET already is able to do that in some cases (Figures 4A, 5, 6), and also emulates the incomplete segmentations by GPET in other cases. A model trained on enhanced "gold standard" segmentations would produce even better segmentations.

Finally, we have focused on segmentation, as it is the most important and most time-consuming step of choroidal analysis. However, the location of the fovea on OCT images must be identified to define the region of interest for derived measurements such as thickness, area, and volume. Identifying the fovea is less time consuming or ambiguous than choroid segmentation, so we plan to extend DeepGPET to

output the fovea location. This would make DeepGPET a fast and efficient end-to-end framework capable of converting a raw OCT image to a set of clinically meaningful segmentation-derived measurements. Likewise, segmenting the choroidal vessels is a very challenging task even for humans and would be prohibitively time consuming to do manually; however, in the future we aim to explore whether DeepGPET can automatically segment the vasculature within the choroid, as well.

Conclusions

Choroid segmentation is a key step in calculating choroidal measurements such as thickness and area. Currently, this is commonly done manually, which is labor intensive and introduces subjectivity. Semiautomatic methods only partially alleviate both of these problems, and previous fully automatic methods have not been easily accessible for researchers. DeepGPET addresses this gap as a fully automatic, end-to-end algorithm that does not require manual interventions. DeepGPET provides performance similar to that of the previously clinically validated, semiautomatic GPET while being fully automatic and an order of magnitude faster. This enables the analysis of large-scale datasets and potential deployment in clinical practice without requiring a trained operator. Although the definition of the lower choroid boundary is still subject to debate (especially when it comes to suprachoroidal spaces), the most important consideration is to have a method that consistently applies the same definition across subjects and studies, which DeepGPET as a fully automatic method does. As an easily accessible, open-source algorithm for choroid segmentation, DeepGPET will enable researchers to easily calculate choroidal measurements much faster and with less subjectivity than before.

Acknowledgments

The authors thank all of the participants in the studies used in this paper, as well as all staff at the Edinburgh Imaging Facility who contributed to image acquisition for this study. We also thank Diana Moukaddem, Niall Strang, Lyle Gray (Glasgow Caledonian University), and Paul McGraw (University of Nottingham) for providing the three Topcon OCT B-scan images.

Supported by a grant from the Medical Research Council (MR/N013166/1 to JB) as part of the Doctoral Training Programme in Precision Medicine at the Usher Institute, University of Edinburgh, and by a grant from UK Research and Innovation (EP/S02431X/1 to JE) as part of the Centre of Doctoral Training in Biomedical AI at the School of Informatics, University of Edinburgh. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) license to any Author Accepted Manuscript version arising.

Disclosure: **J. Burke**, None; **J. Engelmann**, None; **C. Hamid**, None; **M. Reid-Schachter**, None; **T. Pearson**, None; **D. Pugh**, None; **N. Dhaun**, None; **A. Storkey**, None; **S. King**, None; **T.J. MacGillivray**, None; **M.O. Bernabeu**, None; **I.J.C. MacCormick**, None

* JB and JE contributed equally to this work.

References

1. Tan K-A, Gupta P, Agarwal A, et al. State of science: choroidal thickness and systemic health. *Surv Ophthalmol.* 2016;61:566–581.
2. Balmforth C, van Bragt JJMH, Ruijs T, et al. Chorioretinal thinning in chronic kidney disease links to inflammation and endothelial dysfunction. *JCI Insight.* 2016;1:e89173.
3. Robbins CB, Grewal DS, Thompson AC, et al. Choroidal structural analysis in Alzheimer disease, mild cognitive impairment, and cognitively healthy controls. *Am J Ophthalmol.* 2021;223:359–367.
4. Eghtedar RA, Esmaeili M, Peyman A, Akhlaghi M, Rasta SH. An update on choroidal layer segmentation methods in optical coherence tomography images: a review. *J Biomed Phys Eng.* 2022;12:1–20.
5. Masood S, Sheng B, Li P, Shen R, Fang R, Wu Q. Automatic choroid layer segmentation using normalized graph cut. *IET Image Process.* 2018;12:53–59.
6. Salafian B, Kafieh R, Rashno A, Pourazizi M, Sadri S. Automatic segmentation of choroid layer in EDI OCT images using graph theory in neutrosophic space. arXiv. 2018, https://doi.org/10.48550/arXiv.1812.01989.
7. Kajić V, Esmaeelpour M, Považay B, Marshall D, Rosin PL, Drexler W. Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model. *Biomed Ot Express.* 2012;3:86–103.
8. Wang C, Wang YX, Li Y. Automatic choroidal layer segmentation using Markov random field and

level set method. *IEEE J Biomed Health Inform.* 2017;21:1694–1702.
9. Srinath N, Patil A, Kumar VK, Jana S, Chhablani J, Richhariya A. Automated detection of choroid boundary and vessels in optical coherence tomography images. *Annu Int Conf IEEE Eng Med Biol Soc*. 2014;2014:166–169.
10. George N, Jiji CV. Two stage contour evolution for automatic segmentation of choroid and cornea in OCT images. *Biocybern Biomed Eng.* 2019;39:686–696.
11. Danesh H, Kafieh R, Rabbani H, Hajizadeh F. Segmentation of choroidal boundary in enhanced depth imaging OCTs using a multiresolution texture based modeling in graph cuts. *Comput Math Methods Med.* 2014;2014:479268.
12. Chen M, Wang J, Oguz I, VanderBeek BL, Gee JC. Automated segmentation of the choroid in EDI-OCT images with retinal pathology using convolution neural networks. *Fetal Infant Ophthalmic Med Image Anal (2017)*. 2017;10554:177–184.
13. Sui X, Zheng Y, Wei B, et al. Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks. *Neurocomputing.* 2017;237:332–341.
14. Masood S, Fang R, Li P, et al. Automatic choroid layer segmentation from optical coherence tomography images using deep learning. *Sci Rep.* 2019;9:3058.
15. Al-Bander B, Williams BM, Al-Taee MA, Al-Nuaimy W, Zheng Y. A novel choroid segmentation method for retinal diagnosis using deep learning. In: *2017 10th International Conference on Developments in eSystems Engineering (DeSE)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. 2017:182–187.
16. Chen H-J, Huang Y-L, Tse S-L, et al. Application of artificial intelligence and deep learning for choroid segmentation in myopia. *Transl Vis Sci Technol.* 2022;11:38.
17. Zheng G, Jiang Y, Shi C, et al. Deep learning algorithms to segment and quantify the choroidal thickness and vasculature in swept-source optical coherence tomography images. *J Innov Opt Health Sci.* 2021;14:2140002.
18. Engelmann J, Villaplana-Velasco A, Storkey A, Bernabeu MO. Robust and efficient computation of retinal fractal dimension through deep approximation. *arXiv*. 2022, https://doi.org/10.48550/arXiv.2207.05757.
19. Burke J, King S. Edge tracing using Gaussian process regression. *IEEE Trans Image Process.* 2021;31:138–148.
20. Burke J, Pugh D, Farrah T, et al. Evaluation of an automated choroid segmentation algorithm in a longitudinal kidney donor and recipient cohort. *Transl Vis Sci Technol*. 2023;12(11):19, https://doi.org/10.1167/12.11.19.
21. Patterson S. OCT tools. Available at: https://github.com/sarastokes/OCT-tools. Accessed June 20, 2023.
22. Brandt A. OCT marker. Available at: https://github.com/neurodial/OCT-Marker. Accessed June 20, 2023.
23. Kugelman J, Alonso-Caneiro D, Read SA, et al. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Sci Rep.* 2019;9:13298.
24. Xuan M, Wang W, Shi D, et al. A deep learning–based fully automated program for choroidal structure analysis within the region of interest in myopic children. *Transl Vis Sci Technol.* 2023;12:22–22.
25. Mazzaferri J, Beaton L, Hounye G, Sayah DN, Costantino S. Open-source algorithm for automatic choroid segmentation of OCT volume reconstructions. *Sci Rep.* 2017;7:42112.
26. Dhaun N. Optical coherence tomography and nephropathy. The OCTANE study. Available at: https://clinicaltrials.gov/ct2/show/NCT02132741. Accessed November 10, 2023.
27. Pearson T, Chen Y, Dhillon B, Chandran S, van Hemert J, MacGillivray T. Multi-modal retinal scanning to measure retinal thickness and peripheral blood vessels in multiple sclerosis. *Sci Rep.* 2022;12:20472.
28. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *arXiv*. 2015, https://doi.org/10.48550/arXiv.1505.04597.
29. Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3. *arXiv*. 2019, https://doi.org/10.48550/arXiv.1905.02244.
30. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv. 2017, https://doi.org/10.48550/arXiv.1711.05101.
31. Iakubovskii P. Segmentation models PyTorch. Available at: https://github.com/qubvel/segmentation_models.pytorch. Accessed June 10, 2023.
32. Wightman R. PyTorch image models. Available at: https://github.com/rwightman/pytorch-image-models. Accessed June 10, 2023.
33. Early Treatment Diabetic Retinopathy Study Research Group. Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics. ETDRS report number 7. *Ophthalmology.* 1991;98:741–756.

34. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. *J Clin Chem Clin Biochem.* 1983;21:709–720.
35. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. *Perspect Clin Res.* 2017;8:187–191.
36. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;327:307–310.
37. Rahman W, Chen FK, Yeoh J, Patel P, Tufail A, Da Cruz L. Repeatability of manual subfoveal choroidal thickness measurements in healthy subjects using the technique of enhanced depth imaging optical coherence tomography. *Invest Ophthalmol Vis Sci.* 2011;52:2267–2271.
38. Spaide RF, Koizumi H, Pozonni MC. Enhanced depth imaging spectral-domain optical coherence tomography. *Am J Ophthalmol.* 2008;146:496–500.

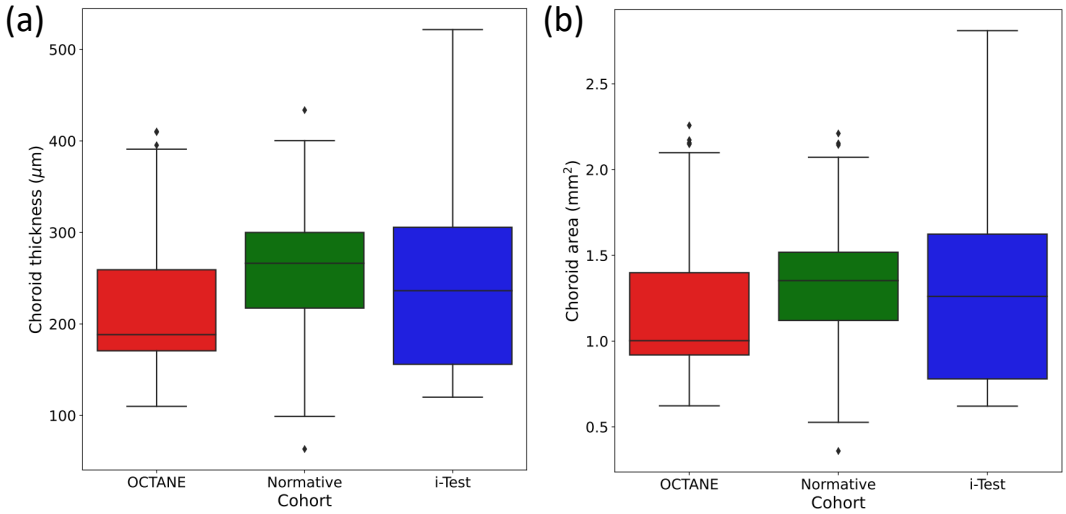


Figure S1: Box-plot distribution plots of choroid thickness (a) and choroid area (b) of the three datasets, OCTANE, i-Test and Normative.

Figure 6.1: Supplementary Figure S1.

Additional supplementary data can be found online: `https://tvst.arvojournals.org/article.aspx?articleid=2793042`.

6.2.3 Second paper

Published under an open license.

Choroidalyzer: An Open-Source, End-to-End Pipeline for Choroidal Analysis in Optical Coherence Tomography

Justin Engelmann,[1,2] Jamie Burke,[3] Charlene Hamid,[4] Megan Reid-Schachter,[4] Dan Pugh,[5] Neeraj Dhaun,[5] Diana Moukaddem,[6] Lyle Gray,[6] Niall Strang,[6] Paul McGraw,[7] Amos Storkey,[8] Paul J. Steptoe,[9] Stuart King,[3] Tom MacGillivray,[4,10] Miguel O. Bernabeu,[2,11] and Ian J. C. MacCormick[8,10]

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2]Centre for Medical Informatics, University of Edinburgh, Edinburgh, United Kingdom
[3]School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom
[4]Clinical Research Facility and Imaging, University of Edinburgh, Edinburgh, United Kingdom
[5]British Heart Foundation Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, United Kingdom
[6]Department of Vision Sciences, Glasgow Caledonian University, Glasgow, United Kingdom
[7]School of Psychology, University of Nottingham, Nottingham, United Kingdom
[8]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[9]Princess Alexandra Eye Pavilion, NHS Lothian, Edinburgh, United Kingdom
[10]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom
[11]The Bayes Centre, University of Edinburgh, Edinburgh, United Kingdom

Correspondence: Justin Engelmann, Nine BioQuarter, Plot 9 Little France Rd., Edinburgh EH16 4UX, UK; Justin.Engelmann@ed.ac.uk.
Jamie Burke, Nine BioQuarter, Plot 9 Little France Rd., Edinburgh EH16 4UX, UK; Jamie.Burke@ed.ac.uk.

JE and JB contributed equally as first authors.

Received: December 5, 2023
Accepted: April 22, 2024
Published: June 4, 2024

Citation: Engelmann J, Burke J, Hamid C, et al. Choroidalyzer: An open-source, end-to-end pipeline for choroidal analysis in optical coherence tomography. *Invest Ophthalmol Vis Sci.* 2024;65(6):6. https://doi.org/10.1167/iovs.65.6.6

Purpose. To develop Choroidalyzer, an open-source, end-to-end pipeline for segmenting the choroid region, vessels, and fovea, and deriving choroidal thickness, area, and vascular index.

Methods. We used 5600 OCT B-scans (233 subjects, six systemic disease cohorts, three device types, two manufacturers). To generate region and vessel ground-truths, we used state-of-the-art automatic methods following manual correction of inaccurate segmentations, with foveal positions manually annotated. We trained a U-Net deep learning model to detect the region, vessels, and fovea to calculate choroid thickness, area, and vascular index in a fovea-centered region of interest. We analyzed segmentation agreement (AUC, Dice) and choroid metrics agreement (Pearson, Spearman, mean absolute error [MAE]) in internal and external test sets. We compared Choroidalyzer to two manual graders on a small subset of external test images and examined cases of high error.

Results. Choroidalyzer took 0.299 seconds per image on a standard laptop and achieved excellent region (Dice: internal 0.9789, external 0.9749), very good vessel segmentation performance (Dice: internal 0.8817, external 0.8703), and excellent fovea location prediction (MAE: internal 3.9 pixels, external 3.4 pixels). For thickness, area, and vascular index, Pearson correlations were 0.9754, 0.9815, and 0.8285 (internal)/0.9831, 0.9779, 0.7948 (external), respectively (all $P < 0.0001$). Choroidalyzer's agreement with graders was comparable to the intergrader agreement across all metrics.

Conclusions. Choroidalyzer is an open-source, end-to-end pipeline that accurately segments the choroid and reliably extracts thickness, area, and vascular index. Especially choroidal vessel segmentation is a difficult and subjective task, and fully automatic methods like Choroidalyzer could provide objectivity and standardization.

Keywords: OCT, choroid, deep learning, automated analysis

The retinal choroid is a densely vascularized tissue at the back of the eye, providing essential nutrients and support to the outer retinal pigment epithelium and photoreceptors.[1] The choroid is emerging as a window into systemic vascular health, including brain,[2] kidney,[3] and heart.[4] The choroid is also affected by ophthalmic conditions like myopia.[5] Thus, the choroid is a potential source of biomarkers for ocular and nonocular disease.[6–9] This is driven by improvements in optical coherence tomography (OCT) imaging, especially enhanced depth imaging OCT (EDI-OCT).[10] Previously, only the retinal layers were well captured, whereas the choroid, which sits below the hyperreflective retinal pigment epithelium, was not imaged well and thus received little attention. Now, the choroid can be captured well and is a promising frontier for systemic health assessment,[11] especially as OCT devices become commonplace even at high street optometrists. To compute choroidal metrics that could serve as potential vascular biomarkers

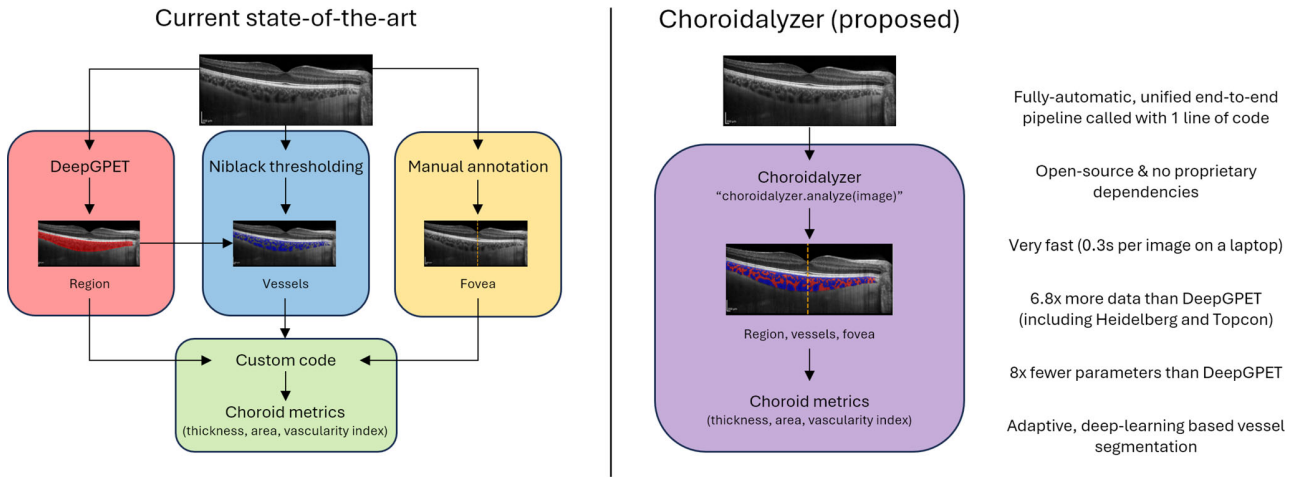


Figure 1. A comparison between Choroidalyzer and the existing state of choroidal analysis. To obtain choroidal metrics in a fovea-centered region of interest, researchers currently need to combine many different tools. Choroidalyzer unifies everything into a end-to-end pipeline that is very fast and convenient to use.

like choroidal thickness, area, or vascular index, the choroid region and vasculature must be identified and segmented accurately and reliably.

While choroidal region segmentation is relatively straightforward compared to vessel segmentation, as only a single shape needs to be identified per scan, accurate detection of the lower choroid boundary (choroid-sclera, C-S, junction) can be time-consuming and at times ambiguous due to poor contrast or image noise. While semiautomatic methods have been proposed,[6,12–20] these typically require training and expertise to use and do not remove human error and subjectivity. Fully automatic, deep learning–based approaches to region segmentation have been proposed and address both the time-intensive and the ambiguous nature of region segmentation, drastically improving both the ease and standardization of choroidal segmentation. Many of these methods are not openly available to the research community,[21–24] but recently DeepGPET, an open-source choroidal region segmentation method, was published that can be freely downloaded from GitHub.[25]

Choroidal vessel segmentation is a far more complex and time-consuming task. The choroidal vessels are highly heterogeneous in terms of vessel size, shape, and edge contrast and are sometimes hard to discern due to poor contrast or noise, making manual segmentations prohibitively time-consuming and very subjective. Currently, local thresholding algorithms are commonplace for choroidal vessel segmentation,[26–28] and the current state-of-the-art is the Niblack algorithm.[29,30] Niblack is a local thresholding technique that segments the vessels using a fixed-size sliding window and a standard deviation offset to determine a pixel-level threshold. However, there is evidence of wide intergrader disagreement between the two commonly used adaptations to Niblack's algorithm.[31] Deep learning approaches have been proposed previously trained on manual annotations or Niblack's algorithm[32,33] but are not openly available at the time of writing.

Finally, in addition to region and vessel segmentation, there are two more necessary steps that are often overlooked, namely, fovea detection and computation of choroidal metrics. OCT B-scans are not necessarily perfectly centered, and the size of a pixel can differ not only between devices but also between scans. Thus, once region, vessels, and fovea are extracted, choroidal metrics should be computed in a fovea-centered region of interest,[6] which must account for key details like the pixel scaling of the scan. Currently, each of these four steps is done by a different tool[34,35] with ad hoc and nonstandardized approaches used especially for fovea detection.[36]

We address these issues by proposing Choroidalyzer, an end-to-end pipeline for choroidal analysis. Choroidalyzer consists of a single deep learning model that simultaneously segments the choroidal region and vessels and detects the fovea location, combined with all the code needed to extract choroidal thickness, area, and vascular index in a fovea-centered region of interest. Figure 1 shows how Choroidalyzer improves on the current state-of-the-art by providing a comprehensive solution for all elements of choroidal analysis. To our knowledge, Choroidalyzer is the first open-source method for comprehensive, automatic analysis of the choroid from a raw OCT B-scan. Choroidalyzer is highly effective, can be run on a standard laptop in less than one-third of a second per image, does not require any specialist training in image processing, and is available on GitHub: https://github.com/justinengelmann/Choroidalyzer.

Methods

Study Population

Our data set contains 5600 OCT B-scans of 233 participants from six cohorts of healthy and diseased individuals, unrelated to ocular pathology: **OCTANE**,[37] a longitudinal cohort study investigating choroidal microvascular changes in renal transplant recipients and healthy donors; **Diurnal Variation**,[37] a subcohort of OCTANE of young individuals investigating the possible effects of diurnal variation on the relationship between the choroid and markers of renal function; **Normative**, a detailed OCT examination of one of the authors (JB) with informed consent; **i-Test**,[37] a cohort of pregnant women evaluating whether the choroidal microvasculature reflects cardiovascular changes in both healthy and complicated pregnancies; **Prevent Dementia**, a longitudinal cohort tracking middle-aged individuals with

Table 1. Overview of Population Characteristics

	OCTANE	Diurnal Variation	Normative	i-Test	Prevent Dementia	GCU Topcon	Total
Subjects	46	20	1	21	121	24	233
Control/case	0/46	20/0	1/0	11/10	56/65	24/0	112/121
Male/female	24/22	11/9	1/0	0/21	66/55	14/9	116/116
Right/left eyes	46/0	20/0	1/1	21/21	117/115	22/21	227/158
Age (mean (SD))	47.5 (12.3)	21.4 (2.3)	23.0 (0.0)	32.8 (5.4)	50.8 (5.6)	21.8 (7.9)	42.9 (13.7)
Device manufacturer	Heidelberg	Heidelberg	Heidelberg	Heidelberg	Heidelberg	Topcon	All
Device type	Standard	Standard	FLEX	FLEX	Standard	DRI Triton Plus	All
Scan location							
Horizontal/vertical	168/0	55/50	4/4	76/76	381/369	132/139	816/638
Volume/radial/peripapillary	0/0/0	0/0/66	365/0/0	2408/0/0	0/0/0	0/1307/0	2773/1307/66
Total B-scans	168	171	373	2560	750	1578	5600

One participant's sex from the GCU Topcon cohort was not recorded.

varying risk of developing late-onset Alzheimer's dementia[38]; and **GCU Topcon**,[39] an investigation into diurnal variation of the choroid in emmetropic and myopic individuals. All studies adhered to the Declaration of Helsinki and received relevant ethical approval, and informed consent from all subjects was obtained in all cases from the host institution. Table 1 describes the population statistics and image acquisition statistics for each cohort.

Three OCT device types were used from two device manufacturers: the spectral domain OCT SPECTRALIS Standard Module OCT1 system and the spectral domain OCT SPECTRALIS portable FLEX Module OCT2 system (both Heidelberg Engineering, Heidelberg, Germany) and the swept source OCT DRI Triton Plus (Topcon, Tokyo, Japan). For the Heidelberg devices, active eye tracking with built-in automatic real time (ART) software was used with horizontal and vertical line scans capturing a 30° (9-mm) fovea-centered region of interest, with an ART of 100 (i.e., each final B-scan is the average of 100 B-scans). Posterior pole macular line scans covered a 30-by-25-degree rectangular region of interest using 31 consecutive scans, each with an ART of 50 (posterior pole scans in the Normative cohort were acquired with an ART of 9). All Heidelberg data were collected at a pixel resolution of 768 × 768 pixels, with a signal quality $\geqslant$15. The Topcon device imaged the macular region using 12 fovea-centered radial scans, spaced 30° apart and covering a 30° (9-mm) region of interest. Each B-scan had a resolution of 992 × 1024 pixels, which was cropped horizontally by 32 pixels and resized to the resolution of the Heidelberg scans of 768 × 768. All Topcon data had an image quality score > 88 determined by the built-in TopQ software.

Five of the six cohorts were split into training (4144 B-scans, 122 subjects), validation (466 B-scans, 28 subjects), and internal test sets (756 B-scans, 37 subjects) containing approximately 75%, 10%, and 15% of the B-scans, respectively. We split the data on the subject level, such that no individual ended up in more than one set. The remaining cohort, OCTANE, was entirely held out as an external test set (168 B-scans, 46 individuals). Supplementary Table S1 gives a detailed overview of population and image characteristics for each of the four sets.

Ground-Truth Labels

The fovea coordinate was defined as the horizontal (column) pixel index, which aligned with the deepest point of the foveal pit depression[36] (i.e., where the central foveal pit was most illuminated, typically aligning with a ridge formed at the photoreceptor layer). The choroidal region was defined as the space posterior to the boundary delineating the retinal pigment epithelium layer and Bruch's membrane complex (RPE-choroid, RPE-C, junction) and superior to the boundary delineating the sclera from the posterior-most point of Haller's layer (C-S junction). Between the choroid and sclera lies the suprachoroidal space, which is rarely visible on OCT B-scans and we consider not to be part of the choroid itself. The choroidal space is made up of interstitial fluid, or stroma, seen as brightly illuminated strips in the OCT B-scans, with interspersed, irregular areas of darker intensity representing choroidal vasculature. This has been both empirically observed[26,40] and widely accepted among the research community.[29] The choriocapillaris, a dense network of choroidal capillaries, is seen as a small band below Bruch's membrane complex approximately 10 microns thick[1] (roughly 3 pixels deep in OCT B-scans) and is assumed as part of the choroidal vasculature alongside larger vessels seen in Haller and Sattler's layers.

For OCT B-scans centered at the fovea (i.e., horizontal, vertical, and radial scans), the foveal column location was detected manually. Those not centered at the fovea do not show the fovea. The ground-truths (GTs) for choroidal region segmentation were generated using DeepGPET[25] with the default threshold of 0.5. In total, 897 scans were excluded from the data set (and removed from Table 1 and Supplementary Table S1) because of poor region segmentations—these were primarily Topcon B-scans that DeepGPET had not been trained on before.

GTs for vessel segmentation were generated using a novel, multiscale quantization and clustering-based approach, called multiscale median cut quantization (MMCQ), which we found to produce superior results to standard application of Niblack in preliminary analysis on the training set. MMCQ segments the choroidal vasculature by performing patchwise local contrast enhancement at several scales using median cut clustering (quantization)[41] and histogram equalization. The pixels of the subsequently enhanced choroidal space are then clustered globally using median cut clustering once more, classifying the pixels belonging to the clusters with the darkest pixel intensities as vasculature. We provide a brief comparison between MMCQ and Niblack in the Supplementary Section 4. In our experience, MMCQ tends to provide higher-fidelity vessel segmentations and avoid oversegmentation compared to Niblack. The code for this algorithm is freely available here at https://github.com/jaburke166/mmcq.

To improve the fidelity and robustness of our vessel segmentation GTs, we randomly varied the brightness and contrast of each OCT B-scan before application of MMCQ. We used five linearly spaced gamma levels to fix the mean brightness of each image between 0.2 and 0.5 and simultaneously altered the contrast using five linearly spaced factors between 0.5 and 3. A 3:2 majority vote for vessel label classification was used across all 25 variants. This improves robustness as spurious over- and undersegmentation contigent on specific image statistics are averaged out.

Choroidalyzer's Deep Learning Model

Choroidalyzer segments the choroid region and vessels, as well as detects the fovea using a UNet deep learning model[42] with a depth of 7. This relatively high depth allows our model to better consider the global context. The first three blocks increase the internal channel dimension from 8 to 64, after which it is kept constant to reduce memory consumption and parameter count. Blocks consist of two convolutional layers, each followed by BatchNorm[43] and ReLU activation. Our up-blocks use a 1×1 convolution to reduce the channel dimension followed by bilinear interpolation, which is more compute and memory efficient than the standard transposed convolutions. We train our model for 40 epochs using the AdamW optimizer[44] with a learning rate of 5×10^{-4} and weight decay of 10^{-8} to minimize binary cross-entropy, clamping the maximum gradient norm to 3 before each step. We use automatic mixed precision to speed up training dramatically while reducing memory consumption by almost half. Forward pass and loss computation are done in bfloat16, a half-precision data type optimized for machine learning.

During training, we apply the following data augmentations in random order per sample: horizontal flip ($P = 0.5$), changing the brightness and contrast independently (factors $\sim U(0.5, 1.5)$, $P = 0.95$), random rotation and shear (degrees $\sim U(-25, 25)$ and $\sim U(-15, 15)$, respectively, $p = 1/3$), and scaling the image (factor $\sim U(0.8, 1.2)$, $p = 1/3$), where $U(a, b)$ denotes a uniform distribution between a and b, and p the probability of the transform being applied. For peripapillary scans that have a resolution of 1536×768, we use a crop of 768×768 using a random multiple of 192 as offset per example and epoch.

The fovea is only a single point that would be difficult for a segmentation model to learn, as predicting close to 0 for all pixels would yield virtually the same loss as a perfect prediction. Thus, we create a target 51 pixels high and 19 pixels wide centered at the GT fovea location. The exact fovea location is set to 1, the whole column to 0.95, and adjacent columns to $0.95 - (d*0.1)$, where d is the column distance from the fovea. Finally, we employ one-sided label smoothing and set all other pixels to 0.01 instead of 0 to stabilize training. We extract fovea column predictions by applying a 21-width triangular filter to the column-wise sums of our model's predictions and taking the column with the highest value.

Statistical Analysis

We evaluate agreement in segmentations using the area under the receiver operating characteristic curve (AUC) and Dice coefficient, applying a fixed threshold of 0.5 to binarize our model's predictions. For the fovea column location, we use mean absolute error (MAE) and median absolute error (AE). For derived choroid metrics, we evaluate agreement with Pearson and Spearman correlations and further report MAEs.

All choroidal metrics were computed using a region of interest (ROI) centered at the foveal pit, measuring 3-mm temporally and nasally—the ROI for volume scans was centered at the middle column index of the image—corresponding to the standardized ROI according to the Early Treatment Diabetic Retinopathy Study (ETDRS) macular grid of 6000×6000 microns.[45] As peripapillary scans do not allow for a fovea-centered region of interest, we only look at segmentation metrics and use a threshold of 0.25 for vessel predictions. Area was computed by counting the pixels within the ETDRS grid, while thickness was measured at three linearly spaced locations, spanning the ETDRS grid, as point-source micron distances between the RPE-C and C-S junctions, locally perpendicular to the RPE-C junction.

Choroid vascular index is the ratio of vessel to nonvessel pixels in the choroid within the ETDRS grid. Our deep learning model outputs probabilities instead of discrete predictions, which capture uncertainty. As capturing uncertainty is desirable, we propose a "soft" vascular index that takes the ratio of predicted probabilities instead of discretized binary predictions. On the validation set, we found that this improves agreement.

To examine and characterize the behavior of our model, we analyzed cases of high error in detail. Concretely, for each of the three tasks (region and vessel segmentation, fovea detection), we selected the 15 examples from each test set where Choroidalyzer produced the highest errors. For redundant cases (i.e., adjacent, highly similar slices from a volume scan), only one was retained. For fovea detection, cases of low error were also discarded. This left 28 cases for region, 29 for vessel, and 25 for fovea.

An adjudicating clinical ophthalmologist (IM) was provided with the original image, Choroidalyzer's prediction, and the GT while being masked to the identity of the methods. Images and labels were provided individually and as composites. For each example, the adjudicator was asked which label they preferred. They also rated each label qualitatively on a five-level ordinal scale ("very bad," "bad," "okay," "good," and "very good") for region segmentation quality, as well as intravascular and interstitial vessel segmentation quality, the latter two to quantify any potential under-segmentation of vessels and oversegmentation of the interstitial space.

Finally, we selected a random subsample of 20 B-scans at the patient level from the external test set to be manually segmented by two graders, M1 and M2. M1 was a clinical ophthalmologist (IM) and M2 was a PhD student who has worked with choroidal OCT data for the last 4 years (JB). Manual graders segmented the region and choroidal vessels using ITK-Snap.[46] The manual segmentations were compared to Choroidalyzer and to the current state-of-the-art, namely, DeepGPET for region segmentations[25] and Niblack for vessel segmentation using a window size of 51 and standard deviation offset of -0.05, which mirrors previously published work.[33]

Results

Performance on Internal and External Test Sets

Table 2 shows the performance of Choroidalyzer on the internal and external test sets. Our model achieves very good performance in terms of AUC and Dice for region and vessels on both sets. Metrics for region are higher than for

Table 2. Metrics for Choroidalyzer Against Ground-Truth Annotations From the Internal and External Test Sets

	Region		Vessel		Fovea		Thickness			Area		MAE	Vascular Index		
Set	AUC	Dice	AUC	Dice	MAE	Median AE	Pearson	Spearman	MAE (µm)	Pearson	Spearman	(mm^2)	Pearson	Spearman	MAE
Internal test	0.9998	0.9789	0.9982	0.8817	3.9 px	3.0 px	0.9754***	0.9692***	8.2252	0.9815***	0.9786***	0.0385	0.8285***	0.8097***	0.0206
External test	0.9998	0.9749	0.9980	0.8703	3.4 px	3.0 px	0.9831***	0.9868***	8.0888	0.9779***	0.9848***	0.0487	0.7948***	0.7991***	0.0306

*** Indicates $P < 0.0001$.

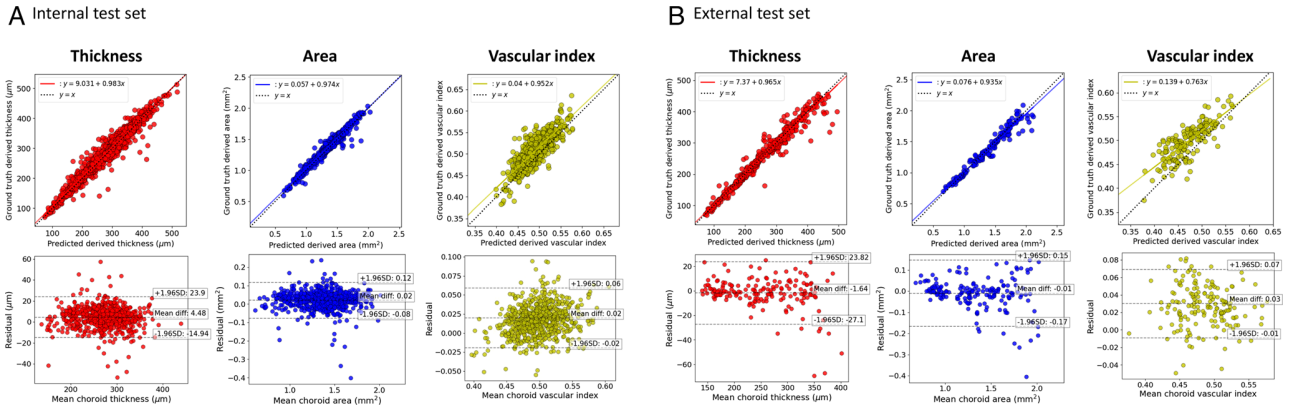


Figure 2. Agreement in thickness, area and vascular index for (A) the internal and (B) the external test sets. *Top row* shows scatterplots with best regression fit and identity lines; *bottom row* shows Bland–Altman plots. Note that we chose to fit each plot to the data range, and thus the scale of the axes is not exactly the same between internal and external test sets, especially for vascular index. Best viewed electronically.

vessels, which is expected as choroidal vessel segmentation is much more difficult and ambiguous than region segmentation, and thus the GTs are themselves imperfect. Performance was slightly higher for the internal test set than the external test set, which is expected, but only marginally so, indicating that our model generalizes well to new cohorts. For the peripapillary scans that only exist in the internal test set, our model achieved an AUC of 0.9996 (region)/0.9925 (vessel) and Dice of 0.9636 (region)/0.7155 (vessel). This is reasonable performance but lower than for other scans.

For fovea detection, the model had a MAE of 3.9 pixels (px) for the internal and 3.4 px for the external test set, with the median absolute error being 3 px for both. This is excellent performance as an error of 3 px on a 768 px-wide image will not meaningfully change our region of interest or resulting metrics (data not shown—see the Supplementary materials for the analysis effects of fovea location on downstream metrics). For the derived choroid metrics, Choroidalyzer shows excellent agreement with the GTs on thickness and area, with Pearson and Spearman correlations of 0.9692 or greater for both internal and external test sets. For the vascular index, performance is a bit lower, with correlations between 0.7948 and 0.8285. Although vascular index depends on both region and vessel segmentation, the other metrics indicate that the differences in vascular index are driven primarily by differences in vessel segmentation. Still, the observed correlations are high in absolute terms. Figure 2 shows correlation and Bland–Altman plots for the three derived metrics on both test sets, which likewise indicate generally very good agreement. Figure 3 shows some examples for each of the three imaging devices.

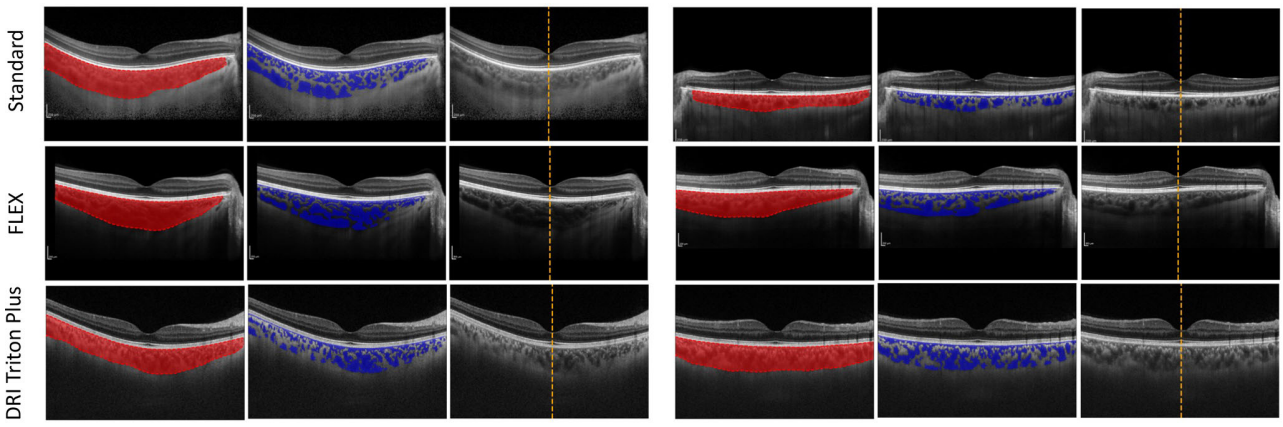


Figure 3. Examples of Choroidalyzer being applied to scans from different imaging devices. Six fovea-centered OCT B-scans, two per imaging device type, from the internal test set showing region segmentations (*left*), vessel segmentations (*middle*), and fovea column location (*right*).

Investigative Ophthalmology & Visual Science

Table 3. Comparison Metrics for the 20 Images Assessed Manually and Algorithmically From the External Test Set

	Region		Vessel		Thickness				Area				Vascular Index			
Comparison	**AUC**	**Dice**	**AUC**	**Dice**	**Pearson**	**Spearman**	**ICC**	**MAE (µm)**	**Pearson**	**Spearman**	**ICC**	**MAE (mm^2)**	**Pearson**	**Spearman**	**ICC**	**MAE**
M1 vs. M2	0.9639	0.9474	0.8891	0.7699	0.9503	0.9521	0.9783	17.8833	0.9516	0.9248	0.9751	0.1096	0.8074	0.6857	0.8172	0.0618
Choroidalyzer vs. Manual (avg)	0.9978	0.9375	0.9914	0.7669	0.9534	0.9663	0.9873	20.9750	0.9554	0.9368	0.9756	0.1150	0.6654	0.7383	0.7613	0.0530
SOTA vs. Manual (avg)	0.9444	0.9333	0.9223	0.7742	0.9676	0.9636	0.9802	19.9250	0.9548	0.9233	0.9769	0.1202	0.6907	0.6105	0.7103	0.1682

Comparisons made between the two manual graders, the proposed model and current state-of-the-art, DeepGPET for region segmentation and the Niblack thresholding algorithm for vessel segmentation. ICC, intraclass correlation; SOTA, current state-of-the-art (i.e., DeepGPET for region and Niblack for vessel segmentation).

Table 4. Mean (Standard Deviation) Execution Time of the Four Different Approaches to Region and Vessel Segmentation for the 20 Images Assessed Manually and Algorithmically From the External Test Set

Method	**Region (s)**	**Vessel (s)**	**Total (s)**
M1	78.400 ± 12.261	1506.000 ± 744.073	1584.400 ± 771.284
M2	165.000 ± 23.889	1176.700 ± 744.073	1341.700 ± 265.800
SOTA	0.751 ± 0.081	0.370 ± 0.105	1.121 ± 0.140
Choroidalyzer	–	–	0.299 ± 0.018

Automated methods were run on a standard laptop with a 4-year-old i5 CPU and 16 GB of RAM but no GPU.

Comparison With Manual Segmentations

Table 3 shows the results from manual segmentations. For automated methods, we compare with each manual grader and then average the performance across both graders to make the results more concise. The comparisons with individual graders are reported in Supplementary Table S2. Interestingly, while vessel Dice for Choroidalyzer (0.7410 vs. M1 and 0.7927 vs. M2; mean 0.7669) is again much worse than region Dice and even worse than the vessel Dice on both test sets, it is very similar to the intergrader agreement of 0.7699. More generally, the intergrader agreements for all other metrics are similar to Choroidalyzer's agreement with the graders, with the notable exception of vascular index. Here, Choroidalyzer's MAE is better (0.0555 vs. M1 and 0.0506 vs. M2; mean 0.0531) than the intergrader agreement (0.0618), as is the Spearman correlation, but Pearson correlation and intraclass correlation are worse. Compared to the respective state-of-the-art (i.e., DeepGPET for region, Niblack for vessel segmentation), Choroidalyzer has better agreement with the graders for most of the metrics, although methods are generally comparable.

Table 4 shows the time per scan for the manual graders and automatic approaches. The manual graders on average needed more than 26 and 22 minutes (mean 24), with the vast majority of that time spent on the vessel segmentation. By contrast, the automatic methods on a standard laptop needed about a second per scan and no human time at all. Thus, to get through a data set of 100 scans, it would take manual graders about 40 hours of work, but with automated methods, it would be less than 2 minutes. With GPU-acceleration, Choroidalyzer and DeepGPET could achieve throughputs of dozens or hundreds of scans per second even on consumer-grade hardware. Comparing the automated methods with each other, Choroidalyzer took 73% less time than DeepGPET and Niblack, while also detecting the fovea location. All three methods are very fast but for very large data sets or deployment on edge devices, Choroidalyzer's efficiency is an additional advantage over existing automated methods.

Detailed Error Analysis

Table 5 shows the results of manual inspection of scans where Choroidalyzer produced the highest error compared to the GT on the test sets. For region segmentation, Choroidalyzer was preferred in 8 cases, the GT in 5, and both methods were considered equally good in 15 cases. In terms of quality, Choroidalyzer was "very bad" in only one case compared with two for the GT and "very good" three times compared to none for the GT. For the vessels, Choroidalyzer was preferred in 13 cases, the GT in 4, and both were tied in 12 cases. Vessel segmentation is a harder task, with no methods achieving "very good." However, the intravascular scores for Choroidalyzer are substantially better, with no "bad" or

Table 5. Preference and Segmentation Scores From Masked Expert Adjudicator (IM) Comparing the Highest Region Segmentation, Vessel Segmentation, and Fovea Column Errors Between Choroidalyzer and the Ground-Truth Labels

	Preferred Choroidalyzer	**Preferred SOTA**	**Both Equally Good**
Region	8/28	5/28	15/28
Vessel	13/29	4/29	12/29
Fovea	23/25	1/25	1/25
Method	**Region: Quality**	**Vessel: Intravascular**	**Vessel: Interstitial**
Choroidalyzer	VG: 3, G: 14, O: 9, B: 1, VB: 1	VG: 0, G: 17, O: 12, B: 0, VB: 0	VG: 0, G: 20, O: 9, B: 0, VB: 0
SOTA	VG: 0, G: 17, O: 8, B: 1, VB: 2	VG: 0, G: 5, O: 19, B: 3, VB: 2	VG: 0, G: 17, O: 8, B: 2, VB: 2

B, bad; G, good; O, okay; VB, very bad; VG, very good.

Investigative Ophthalmology & Visual Science

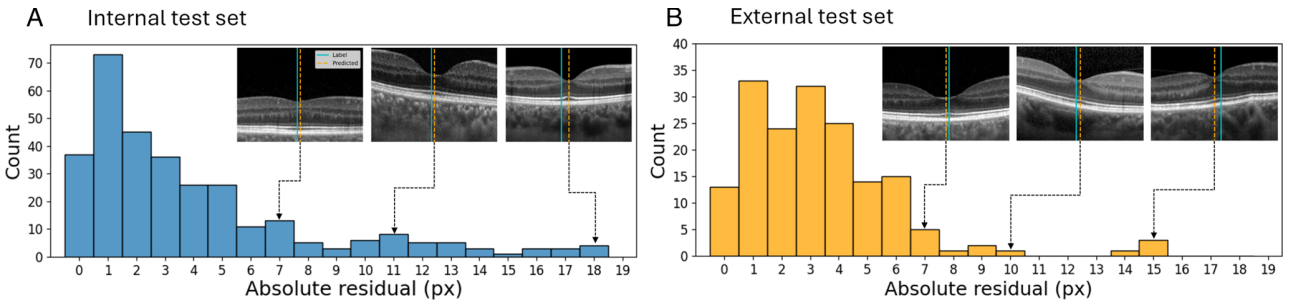


Figure 4. Histogram of absolute errors for fovea column detection for the internal (*left*) and external (*right*) test sets. Examples for different levels of error are shown, with *dotted lines* indicating which part of the distribution they come from. In the examples, the *teal line* indicates the GT label, the *dashed orange line* the prediction.

"very bad" (vs. 3 and 2, respectively for GT) and far more "good" (17 vs. 5), and the interstitial scores are similarly better. Finally, for the fovea, Choroidalyzer was preferred 23 of 25 times and the GT only twice, indicating that large fovea errors are almost exclusively due to mistakes in the manual GT labels.

Figure 4 shows the distributions of fovea errors for both test sets along with each example in both sets. For very large residuals (10+ px), the GTs are wrong and Choroidalyzer correctly identifies the fovea location. For errors around 7 px, still twice the MAE, both methods are similar, with either method sometimes being more correct. Further exploration revealed the majority of incorrectly labeled ground-truths to be Topcon OCT B-scans, as each 12-stack of radial scans are not centered at the fovea, and initial manual annotation detected the fovea for only one to represent each stack. Despite this oversight, Choroidalyzer learned to dectect the fovea robust and accurately.

Discussion

We developed Choroidalyzer, an end-to-end pipeline for choroidal analysis. Choroidalyzer shows excellent performance on the internal and external test sets. Choroidalyzer produced the highest errors, primarily cases of imperfect GTs, and Choroidalyzer was generally preferred by a blinded adjudicating ophthalmologist (IM), further indicating robustness and good performance. Its agreement with manual segmentations, which demand substantial time and attention from a human expert, is comparable to the intergrader agreement. This suggests that Choroidalyzer performs well compared to laborious manual segmentation and also highlights the subjectivity introduced by manual graders. Choroidalyzer not only produces results similar to that of a skilled manual grader but also does so fully automatically without introducing subjectivity and thus increases standardization and reproducibility. If researchers use Choroidalyzer, their results are repeatable and would be much more comparable to other studies also using Choroidalyzer than if different manual graders were used in each case.

Additionally, Choroidalyzer saves a substantial amount of time per image over manual segmentation, freeing up researcher time and enabling large-scale analyses that otherwise would not be possible. Even compared to the current state-of-the-art for automated methods, DeepGPET and Niblack, Choroidalyzer can do the analysis in roughly a quarter of the time. More importantly, Choroidalyzer provides an end-to-end pipeline, which makes it easier to implement and use than having to combine multiple methods like Niblack and DeepGPET. Ease of use is often underappreciated in the literature but key in saving researchers time and allowing them to focus on the science.

Choroidalyzer performed well against manual graders relative to the state-of-the-art methods, reaching or surpassing the levels of agreement even between the two manual graders, particularly for vascular index, a far more difficult metric to calculate accurately than area and thickness. The intergrader agreement between manual graders for these metrics indicates a potential lower bound of what effect sizes we might expect from these metrics. This has important downstream impact on the statistical confidence of results from cohort studies, particularly when assessing the choroidal vasculature.[28,31]

It is often difficult to visualize the choroid due to imaging noise, poor eye tracking and patient fixation, or operator inexperience. Thus, in some cases, vessel boundaries can be hard to discern. This is why we proposed to use a soft version of the choroid vascular index, where the probabilities that Choroidalyzer outputs are used instead of thresholded, binarized segmentations. The probabilities capture uncertainty about the precise location of the vessel wall and thus are more robust than using a single, somewhat arbitrary threshold. Users could also tune the binarization threshold for their own images, if desired, which might help in instances of poor visibility of the choroidal vasculature.

Segmentation performance for peripapillary scans was reasonable but much worse than for other scan types. This could be due to those scans being relatively rare in our data set and showing parts of the retina on the nasal side of the optic disc that are not captured in fovea-centered scans. More peripapillary training data would likely increase performance. In our opinion, at present, Choroidalyzer can be used for these scans but requires subsequent manual inspection and potential correction. Furthermore, adjusting the binarization threshold for the vessel predictions can improve results.

Our model detected the fovea well, and the largest errors were cases where ground-truths were incorrectly labeled with the model correctly identifying the fovea location as confirmed by masked adjudication. Thus, the model performed even better than what the quantitative results suggest. In the present work, we have focused on identifying the fovea column, which is needed to define the fovea-centered region of interest. However, after selecting and evaluating our final model, we realized that in relatively

Investigative Ophthalmology & Visual Science

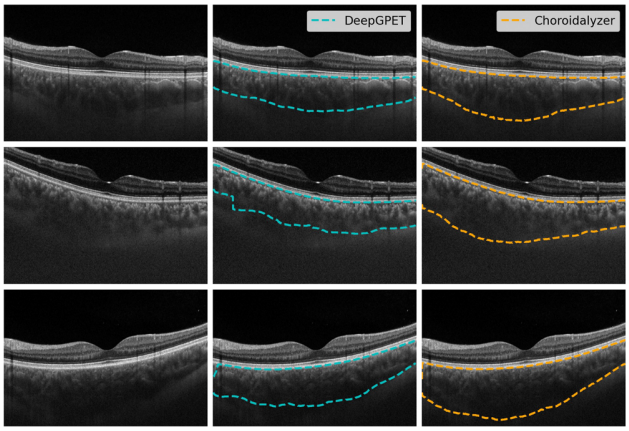


Figure 5. Three example Topcon OCT B-scans with successful region segmentations from Choroidalyzer (*right*) and failed segmentations from DeepGPET (*middle*).

rare cases related to poor image acquisition, the retina and choroid can be at a steep angle relative to the image axes. For those, it would be best to define the region of interest along the choroid axes rather than image axes, most easily done by drawing a center line from the fovea perpendicularly through the retina and choroid. Thus, it could be useful to also segment the retina and to determine both the row and column of the fovea. While not our initial objective, we did some preliminary analyses and found that we can derive the fovea row well with our current model (data not shown). We also analyzed the effect of defining the region of interest perpendicular to the choroid instead of aligned to the image axes, as shown in Supplementary Section 5. Contrary to our initial hypothesis, the difference in area and vascular index is very minor even for highly myopic eyes. However, for thickness, the choroid-aligned measurement tends to be higher, and there are a few cases of large disagreement. Furthermore, to understand the effect of fovea location error on downstream choroidal metrics, we simulated random per-sample deviations of ±6 px, twice the median AE, and found that they yielded virtually identical results (see Supplementary Fig. S1 and Fig. S2).

The data set in the present work was substantially larger than the one used for DeepGPET and importantly contains both Heidelberg and Topcon scans. As a result, Choroidalyzer can segment even difficult Topcon scans where DeepGPET failed (Fig. 5). Choroidalyzer was trained on region and vessel GTs generated by fully and semiautomatic methods, respectively, which were then checked for errors and only manually improved where needed. Recent work argues that such approaches to generating GTs are preferable as they reduce subjectivity and thus bias and inconsistency.[47]

Choroidalyzer also has limitations. Most importantly, there is no quality scoring component to reject B-scans that do not show the choroid in sufficient detail to allow for reasonable analysis. While modern OCT devices typically show the choroid in good detail, especially if EDI is used, this is not always the case. Most devices provide some quality indicators, but we have not investigated quality thresholds for specific devices, below which Choroidalyzer would not function. Furthermore, OCT quality indicators are typically focused on the retina, and although poor visualization of the retina might imply poor visualization of the choroid, the reverse is not necessarily the case. A quality scoring method specific to the choroid would be a useful addition to the field. Another limitation is that Choroidalyzer was trained only on cohorts relating to systemic health but not ocular disease or data acquired during routine clinical practice.

Future work could improve the underlying deep learning model of Choroidalyzer (e.g., by training and evaluating it on data from more diverse sources). Data with ocular pathology (e.g., abnormally sized choroids due to myopia, age-related macular degeneration, or central serous chorio-retinopathy) could be used to investigate whether Choroidalyzer is robust in those contexts and to train an improved version if needed. Moreover, automated quality scoring methods relating to the choroid would address a key need in choroidal analysis. Finally, Choroidalyzer could be extended to measure additional choroidal metrics, such as macular thickness and vessel density maps across a volume, or relating to its curvature.

Conclusions

Choroidal thickness, area, and especially vascular index are highly interesting metrics and potential biomarkers for both systemic and ocular health. However, calculating them used to be laborious and—when done manually—subjective. Choroidalyzer provides an efficient, end-to-end pipeline to alleviate these problems. We hope that by making Choroidalyzer openly accessible, we will enable researchers and clinicians to conveniently calculate these metrics and use them for their research, while improving reproducibility and standardization in the field.

Acknowledgments

The authors thank the Edinburgh Imaging and Edinburgh Clinical Research Facility at the University of Edinburgh for support and all participants in the studies used in this article. Supported in part by the Alzheimer's Drug Discovery Foundation (project no. GDAPB-201808-2016196), NHS Lothian R&D, and British Heart Foundation Centre for Research Excellence Award III (RE/18/5/34216). Supported in part also by the Wellcome Leap In Utero scheme. The funding sources were not involved in designing, conducting, or submitting this work.

M.O.B. gratefully acknowledges funding from: Fondation Leducq Transatlantic Network of Excellence (17 CVD 03); EPSRC grant no. EP/X025705/1; British Heart Foundation and The Alan Turing Institute Cardiovascular Data Science Award (C-10180357); Diabetes UK (20/0006221); Fight for Sight (5137/5138); the SCONe projects funded by Chief Scientist Office, Edinburgh & Lothians Health Foundation, Sight Scotland, the Royal College of Surgeons of Edinburgh, the RS Macdonald Charitable Trust, and Fight For Sight.

Supported by UK Research and Innovation (grant EP/S02431X/1) as part of the Centre of Doctoral Training in Biomedical AI at the School of Informatics, University of Edinburgh (JE). Supported by the Medical Research Council (grant MR/N013166/1) as part of the Doctoral Training Programme in Precision Medicine at the Usher Institute, University of Edinburgh (JB).

Disclosure: **J. Engelmann**, None; **J. Burke**, None; **C. Hamid**, None; **M. Reid-Schachter**, None; **D. Pugh**, None; **N. Dhaun**, None; **D. Moukaddem**, None; **L. Gray**, None; **N. Strang**, None; **P. McGraw**, None; **A. Storkey**, None; **P.J. Steptoe**, None; **S. King**, None; **T. MacGillivray**, None; **M.O. Bernabeu**, None; **I.J.C. MacCormick**, None

References

1. Nickla DL, Wallman J. The multifunctional choroid. *Prog Retin Eye Res*. 2010;29(2):144–168.
2. Robbins CB, Grewal DS, Thompson AC, et al. Choroidal structural analysis in Alzheimer disease, mild cognitive impairment, and cognitively healthy controls. *Am J Ophthalmol*. 2021;223:359–367.
3. Balmforth C, van Bragt JJ, Ruijs T, et al. Chorioretinal thinning in chronic kidney disease links to inflammation and endothelial dysfunction. *JCI Insight*. 2016;1(20):e89173, https://doi.org/10.1172/jci.insight.89173.
4. Yeung SC, You Y, Howe KL, Yan P. Choroidal thickness in patients with cardiovascular disease: a review. *Surv Ophthalmol*. 2020;65(4):473–486.
5. Read SA, Fuss JA, Vincent SJ, Collins MJ, Alonso-Caneiro D. Choroidal changes in human myopia: insights from optical coherence tomography imaging. *Clin Exp Optom*. 2019;102(3):270–285.
6. Burke J, Pugh D, Farrah T, et al. Evaluation of an automated choroid segmentation algorithm in a longitudinal kidney donor and recipient cohort. *Transl Vis Sci Technol*. 2023;12(11):19.
7. Burke J, Dhaun N, Dhillon B, Wilson KJ, Beare NAV, MacCormick IJC. The retinal contribution to the kidney–brain axis in severe malaria. *Trends Parasitol*. 2023;39(6):410–411, ISSN 1471–4922, https://doi.org/10.1016/j.pt.2023.03.002.
8. Shin YU, Lee SE, Kang MHO, Han S-W, Yi J-H, Cho H. Evaluation of changes in choroidal thickness and the choroidal vascularity index after hemodialysis in patients with end-stage renal disease by using swept-source optical coherence tomography. *Medicine (Baltimore)*. 2019;98(18).
9. Kundu A, Ma JP, Robbins CB, et al. Longitudinal analysis of retinal microvascular and choroidal imaging parameters in Parkinson's disease compared with controls. *Ophthalmol Sci*. 2023;3(4):100393, ISSN 2666–9145, https://doi.org/10.1016/j.xops.2023.100393.
10. Spaide RF, Koizumi H, Pozonni MC. Enhanced depth imaging spectral-domain optical coherence tomography. *Am J Ophthalmol*. 2008;146(4):496–500.
11. Tan K-A, Gupta P, Agarwal A, et al. State of science: choroidal thickness and systemic health. *Surv Ophthalmol*. 2016;61(5):566–581.
12. Burke J, King S. Edge tracing using gaussian process regression. *IEEE Trans Image Process*. 2021;31:138–148.
13. Eghtedar RA, Esmaeili M, Peyman A, Akhlaghi M, Rasta SH. An update on choroidal layer segmentation methods in optical coherence tomography images: a review. *J Biomed Phys Eng*. 2022;12(1):1.
14. Masood S, Sheng B, Li P, Shen R, Fang R, Wu Q. Automatic choroid layer segmentation using normalized graph cut. *IET Image Proc*. 2018;12(1):53–59.
15. Salafian B, Kafieh R, Rashno A, Pourazizi M, Sadri S. Automatic segmentation of choroid layer in EDI OCT images using graph theory in neutrosophic space. arXiv preprint arXiv:1812.01989, 2018.
16. Kajić V, Esmaeelpour M, Považay B, Marshall D, Rosin PL, Drexler W. Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model. *Biomed Opt Express*. 2012;3(1):86–103.
17. Wang C, Wang YaX, Li Y. Automatic choroidal layer segmentation using Markov random field and level set method. *IEEE J Biomed Health Inf*. 2017;21(6):1694–1702.
18. Srinath N, Patil A, Kumar VK, Jana S, Chhablani J, Richhariya A. Automated detection of choroid boundary and vessels in optical coherence tomography images. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Chicago, IL, USA: IEEE; 2014:166–169, doi: 10.1109/EMBC.2014.6943555.
19. George N, Jiji CV. Two stage contour evolution for automatic segmentation of choroid and cornea in OCT images. *Biocybern Biomed Eng*. 2019;39(3):686–696.
20. Danesh H, Kafieh R, Rabbani H, Hajizadeh F. Segmentation of choroidal boundary in enhanced depth imaging octs using a multiresolution texture based modeling in graph cuts. *Comput Math Methods Med*. 2014;2014:9, https://doi.org/10.1155/2014/479268.
21. Mazzaferri J, Beaton L, Hounye G, Sayah DN, Costantino S. Open-source algorithm for automatic choroid segmentation of OCT volume reconstructions. *Sci Rep*. 2017;7(1):42112.
22. Kugelman J, Alonso-Caneiro D, Read SA, et al. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Sci Rep*. 2019;9(1):13298.
23. Devalla SK, Renukanand PK, Sreedhar B-K, et al. Drunet: a dilated-residual U-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomed Opt Express*. 2018;9(7):3244–3265.
24. Chen H-J, Huang Y-L, Tse S-L, et al. Application of artificial intelligence and deep learning for choroid segmentation in myopia. *Transl Vis Sci Technol*. 2022;11(2):38–38.
25. Burke J, Engelmann J, Hamid C, et al. An open-source deep learning algorithmfor efficient and fully-automatic analysis of the choroid in optical coherence tomography. *Trans. Vis. Sci. Tech*. 2023;12(11):27, https://doi.org/10.1167/tvst.12.11.27.
26. Branchini LA, Adhi M, Regatieri CV, et al. Analysis of choroidal morphologic features and vasculature in healthy eyes using spectral-domain optical coherence tomography. *Ophthalmology*. 2013;120(9):1901–1908.
27. Sonoda S, Sakamoto T, Yamashita T, et al. Choroidal structure in normal eyes and after photodynamic therapy determined by binarization of optical coherence tomographic images. *Invest Ophthalmol Vis Sci*. 2014;55(6):3893–3899.
28. Agrawal R, Gupta P, Tan K-A, et al. Choroidal vascularity index as a measure of vascular status of the choroid: measurements in healthy eyes from a population-based study. *Sci Rep*. 2016;6(1):21090.
29. Agrawal R, Ding J, Sen P, et al. Exploring choroidal angioarchitecture in health and disease using choroidal vascularity index. *Prog Retin Eye Res*. 2020;77:100829.
30. Betzler BK, Ding J, Wei X, et al. Choroidal vascularity index: a step towards software as a medical device. *Br J Ophthalmol*. 2022;106(2):149–155.
31. Wei X, Sonoda S, Mishra C, et al. Comparison of choroidal vascularity markers on optical coherence tomography using two-image binarization techniques. *Invest Ophthalmol Vis Sci*. 2018;59(3):1206–1211.
32. Liu X, Bi L, Xu Y, Feng D, Kim J, Xu X. Robust deep learning method for choroidal vessel segmentation on swept source optical coherence tomography images. *Biomed Opt Express*. 2019;10(4):1601–1612.
33. Muller J, Alonso-Caneiro D, Read SA, Vincent SJ, Collins MJ. Application of deep learning methods for binarization of the choroid in optical coherence tomography images. *Transl Vis Sci Technol*. 2022;11(2):23.
34. Zheng Gu, Jiang Y, Shi Ce, et al. Deep learning algorithms to segment and quantify the choroidal thickness and vasculature in swept-source optical coherence tomography images. *J Innov Opt Health Sci*. 2021;14(1):2140002.
35. Khaing TT, Okamoto T, Ye C, et al. Choroidnet: a dense dilated U-net model for choroid layer and vessel segmentation in optical coherence tomography images. *IEEE Access*. 2021;9:150951–150965.
36. Xuan M, Wang W, Shi D, et al. A deep learning–based fully automated program for choroidal structure analysis within

the region of interest in myopic children. *Transl Vis Sci Technol*. 2023;12(3):22.
37. Dhaun N. Optical coherence tomography and nephropathy: the Octane Study. 2014. https://clinicaltrials.gov/ct2/show/NCT02132741. Accessed May 31, 2023.
38. Ritchie CW, Ritchie K. The prevent study: a prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer's disease. *BMJ Open*. 2012;2(6):e001893.
39. Moukaddem D, Strang N, Gray L, McGraw P, Scholes C. Comparison of diurnal variations in ocular biometrics and intraocular pressure between hyperopes and non-hyperopes. *Invest Ophthalmol Vis Sci*. 2022;63(7):1428–F0386.
40. Sohrab M, Wu K, Fawzi AA. A pilot study of morphometric analysis of choroidal vasculature in vivo, using en face optical coherence tomography. *PLoS One*. 2012;7(11):e48631.
41. Heckbert P. Color image quantization for frame buffer display. *ACM SIGGRAPH Comput Graph*. 1982;16(3):297–307.
42. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. (eds) *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*. Springer; 2015;9351:234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
43. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. 2015;37:448–456, (ICML'15), JMLR.org.
44. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
45. Early Treatment Diabetic Retinopathy Study Research Group. Early treatment diabetic retinopathy study design and baseline patient characteristics: ETDRS Report Number 7. *Ophthalmology*. 1991;98(5):741–756.
46. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116–1128.
47. Maloca PM, Pfau M, Janeschitz-Kriegl L, et al. Human selection bias drives the linear nature of the more ground truth effect in explainable deep learning optical coherence tomography image segmentation. *J Biophotonics*. 2024;17(2):e202300274, https://doi.org/10.1002/jbio.202300274.
48. Rahman W, Chen FK, Yeoh J, Patel P, Tufail A, Da Cruz L. Repeatability of manual subfoveal choroidal thickness measurements in healthy subjects using the technique of enhanced depth imaging optical coherence tomography. *Invest Ophthalmol Vis Sci*. 2011;52(5):2267–2271.
49. Agrawal R, Wei X, Goud A, Vupparaboina KK, Jana S, Chhablani J. Influence of scanning area on choroidal vascularity index measurement using optical coherence tomography. *Acta Ophthalmol (Copenh)*. 2017;95(8):e770–e775.

Investigative Ophthalmology & Visual Science

Supplementary Material

1. Population statistics across training, validation and test sets

	Training	Validation	Testing	External test	Total
Subjects	122	28	37	46	233
Male/Female	64 / 57	12 / 16	16 / 21	24 / 22	116 / 116
Control/Case	76 / 46	16 / 12	20 / 17	0 / 46	112 / 121
Right/Left eyes	117 / 107	27 / 23	37 / 28	46 / 0	227 / 158
Standard/FLEX/DRI Triton Plus	88 / 14 / 20	24 / 2 / 2	29 / 6 / 2	46 / 0 / 0	187 / 22 / 24
Heidelberg/Topcon	102 / 20	26 / 2	35 / 2	46 / 0	209 / 24
Age (mean (SD))	40.7 (14.2)	42.5 (11.9)	44.5 (13.4)	47.5 (12.3)	42.9 (13.4)
Cohort					
OCTANE	0	0	0	46	46
Diurnal Variation	12	4	4	0	20
Normative	1	0	0	0	1
i-Test	13	2	6	0	21
Prevent Dementia	76	20	25	0	121
GCU Topcon	20	2	2	0	24
B-scans					
Standard/Flex/DRI Triton Plus	582 / 2,281 / 1,281	136 / 190 / 140	137 / 462 / 157	168 / 0 / 0	1,023 / 2,933 / 1,578
Heidelberg/TopCon	2,863 / 1,281	326 / 140	599 / 157	168 / 0	3,956 / 1,578
Horizontal/Vertical scans	462 / 461	90 / 82	95 / 95	168 / 0	816 / 638
Volume/Radial/Peripapilary scans	2,161 / 1,060 / 39	178 / 116 / 15	434 / 131 / 12	0 / 0 / 0	2,773 1,307 / 0
Total B-scans	4,183	481	768	168	5,600

Table S1 Overview of population and image characteristics of the internal training, validation and test sets, and also the external test set. Note that one participant's sex from the Topcon cohort was not recorded. SD: Standard Deviation.

2. Full comparison metrics with methods and manual graders

Comparison	Region		Vessel		Thickness				Area				Vascular Index			
	AUC	Dice	AUC	Dice	pearson	spearman	ICC	MAE (μm)	pearson	spearman	ICC	MAE (mm^2)	pearson	spearman	ICC	MAE
M1 vs. M2	0.9639	0.9474	0.8891	0.7699	0.9503	0.9521	0.9783	17.8833	0.9516	0.9248	0.9751	0.1096	0.8074	0.6857	0.8172	0.0618
M1																
Choroidalyzer	**0.9964**	**0.9242**	**0.9896**	0.7410	0.9322	**0.9490**	**0.9761**	27.2167	**0.9211**	**0.8872**	0.9570	**0.1598**	**0.7668**	**0.8406**	**0.7265**	**0.0555**
SOTA	0.9370	0.9227	0.9271	**0.7714**	**0.9437**	0.9378	0.9676	**25.8500**	0.9198	0.8692	**0.9589**	0.1631	0.7150	0.6857	0.7157	0.1901
M2																
Choroidalyzer	**0.9993**	**0.9507**	**0.9933**	**0.7927**	0.9746	0.9838	**0.9984**	14.7333	0.9896	**0.9865**	0.9942	**0.0702**	0.5640	**0.6361**	**0.7960**	**0.0506**
SOTA	0.9175	0.9439	0.9175	0.7770	**0.9914**	**0.9894**	0.9927	**14.0000**	**0.9897**	0.9774	**0.9948**	0.0770	**0.6663**	0.5353	0.7047	0.1464

Table S2 Full comparison metrics between Choroidalyzer, two manual graders M1 and M2, and state of the art region and vessel segmentation methods DeepGPET and Niblack. AUC: Area under the Receiver Operating Characteristic Curve. MAE: Mean Absolute Error. ICC: Intra-Class Correlation.

3. Analysis effects of fovea location error on downstream metrics

Choroidalyzer measured the fovea column coordinate with a median absolute error of 3 pixels in both the internal and external test sets. We tested the effect of perturbing the fovea column on choroidal metrics by comparing fovea-centred metrics and metrics derived after the fovea column was randomly perturbed using a discretised uniform distribution $\sim U(-6, 6\}$ (excluding 0). 50 simulations were run on approximately 10% of the dataset (495 OCT B-scans), selected at random to represent eye type and location on the macula (see supplementary Table S3 for a description on the image statistics of this random sample).

All metrics had excellent Pearson correlation ($r > 0.99$, p<0.00001, supplementary Fig. S1). Scatterplots of metrics for the poorest performing simulation according to absolute error across all metrics (supplementary Fig. S2) shows excellent agreement with the

Eyes (Number of scans)	OD	OS		Total
	42 (263)	31 (232)		73 (495)
Location	H-line/V-line	Ppole/Radial		
	85/64	217/129		495
Device	OCT1 (Heidelberg)	OCT2 (Heidelberg)	DRI Triton Plus (Topcon)	
	113	225	157	495

Table S3 Image statistics of the random sample of 495 OCT B-scans used to understand the effects of random perturbations of the fovea coordinate.

identity line, with limits of agreement in the Bland-Altman plots well within acceptable bounds for all metrics[48;49]. Thus, the fovea column quantitative error observed from Choroidalyzer does not significantly impact the choroidal metrics.

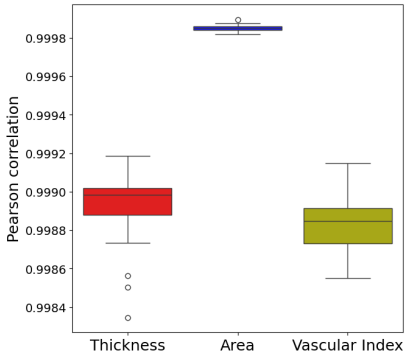


Figure S1 Distribution of Pearson correlation coefficients for each choroidal metric when perturbing the fovea coordinate column. Note the scale of the y-axis, even the lowest correlation we observed was > 0.99.

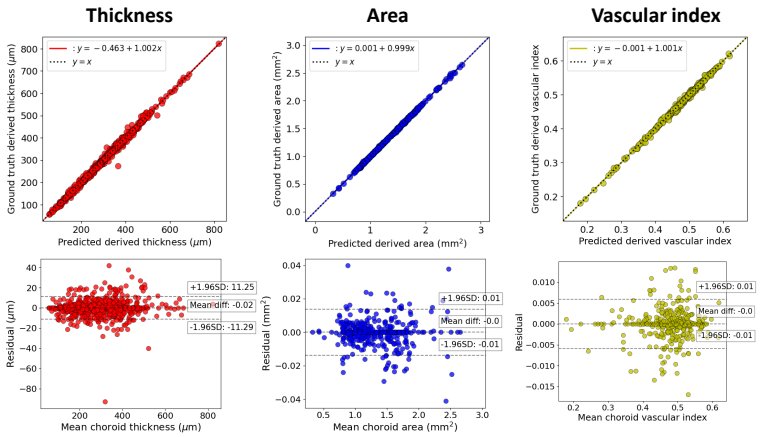


Figure S2 Correlation and Bland-Altman plots for choroidal metrics for the poorest performing simulation of perturbing the fovea column coordinate on a random 10% subsample of the dataset.

4. Comparison of MMCQ and Niblack segmentation methods

Fig. S3 shows some qualitative examples between the vessel segmentations produced by MMCQ and Niblack for exemplar B-scans from all imaging devices used in this study. Table S4 shows the results of comparing MMCQ and Niblack with the 20 OCT B-scans from the external test set which were manually segmented by two experienced graders (I.M. and J.B.). We found that both approaches performed similarly when compared to the manual grader. We did observe a large mean absolute error in CVI when comparing the two approaches directly. We believe this is due to Niblack having a tendency to oversegment the choroid, such that it is able to segment all vessels in the choroid at the cost of segmenting parts of the interstitial space. MMCQ instead attempts to preserve vessel fidelity by not segmenting the interstitial space — at the cost of rejecting ambiguous pixels which either represent vessel walls or interstitial space.

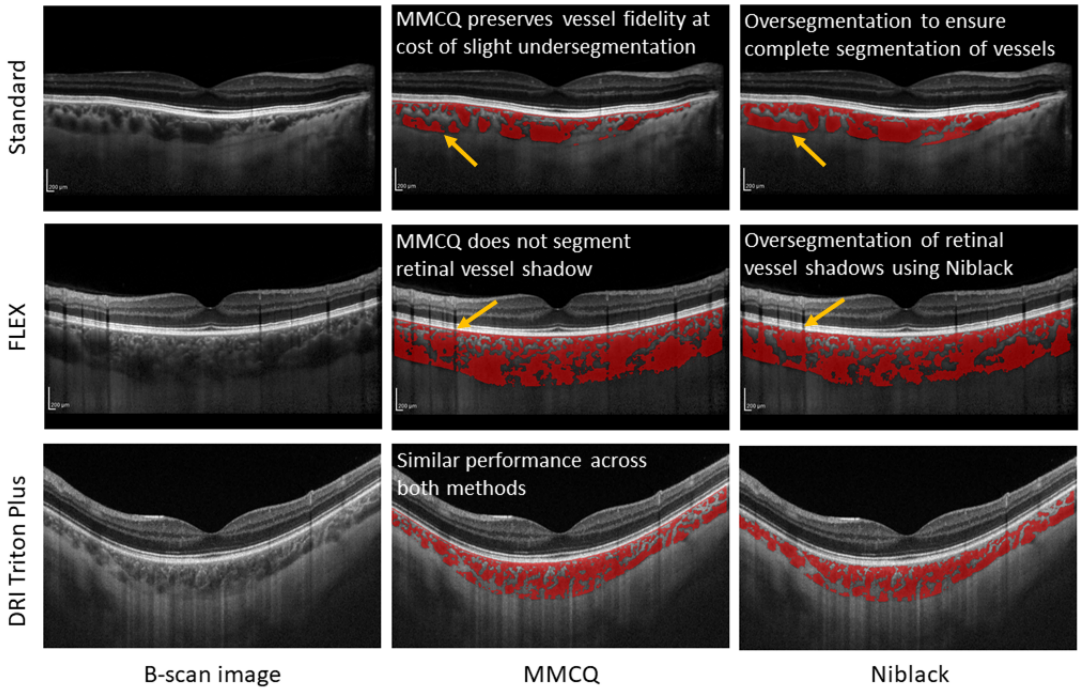


Figure S3 Examples of MMCQ (centre column) and Niblack (right-hand column) segmenting exemplar OCT B-scans from each imaging device, the Heidelberg Standard module, the Heidelberg FLEX module, and the Topcon DRI Trition plus.

Comparison	CVI			Vessel segmentation	
	Pearson	Spearman	MAE	Dice	AUC
Niblack vs. Manual (Avg)	0.560916	0.663158	**0.083508**	0.746164	0.820954
MMCQ vs. Manual (Avg)	**0.679932**	0.627068	0.088144	**0.819151**	**0.903037**
Niblack vs. MMCQ	0.699646	0.781955	0.159811	0.777819	0.948477

Table S4 Vessel segmentation and choroid-derived CVI metrics between average manual segmentation, Niblack thresholding algorithm and MMCQ. CVI: Choroid vascular index.

5. Analysis effects of Choroid- and Image-aligned regions of interest

We investigated the effects that different regions of interest aligned with either image axis or choroid axis had on choroid measurements. Our initial hypothesis was that B-scans of highly myopic eyes could skew the choroid off-centre from the image axis, which could have a noticeable impact on choroidal measurements. We conducted two forms of analysis to test this hypothesis, in both cases comparing measurements made according to the horizontal image axis, and the choroid axes.

First, we used the same random sample of OCT B-scans used in supplementary section 3 (Table S3). We compared choroidal measures for different axis alignment (Fig. S4) and found that choroidal thickness differed significantly between different alignments as the size of the choroid increased, while area and CVI remained highly reproducible. We suspect this is because choroid thickness is

a one-dimensional straight line distance measure, which can be highly susceptible to changes in pixel length-scale. This is emphasised in OCT B-scans as the axial and lateral resolution are different (approximately 4:1) in Heidelberg and Topcon imaging devices.

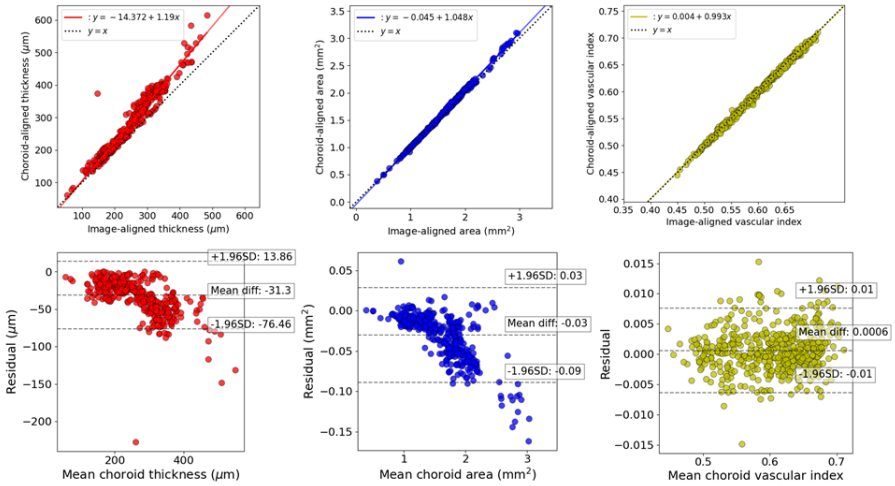


Figure S4 Correlation and bland-Altman plots of measuring thickness, area and CVI using the choroid axis (x) and the image axis (y).

Secondly, we investigated any effect that myopia might have on these measurements. We selected three individuals from the GCU Topcon cohort: one highly hypermetropic (6.375 Spherical Equivalent Dioptres (D)), one emmetropic (0.75D) and one highly myopic (-7.5D) individual. A random B-scan was segmented and thickness (at macular locations described in the main paper, i.e. subfoveal and 2mm temporal and nasal to the fovea), area and CVI was measured. Table S5 shows the measurements for these three OCT B-scans, measuring both aligned to the choroid axis and the image axis. Fig. S5 shows the three OCT B-scans with choroidal thickness and area measurements annotated.

Individual	Thickness, (N, F, T) (microns)			Area (mm2)			Choroid vascular index (CVI)		
	Choroid	Image	Largest residual	Choroid	Image	Residual	Choroid	Image	Residual
Hypermetropic (6.375D)	(541, 762, 442)	(452, 686, 355)	89 microns (N)	3.481	3.497	0.015	0.731	0.728	0.003
Emmetropic (0.75D)	(248, 377, 427)	(232, 337, 318)	109 microns (T)	1.928	1.881	0.046	0.607	0.606	0.001
Myopic (-7.5D)	(212, 391, 474))	(198, 342, 368)	106 microns (T)	1.928	1.880	0.048	0.697	0.692	0.005

Table S5 Comparisons of thickness, area and vascular index measured aligned with the choroid axis and with the image axis. For each choroid measure, the values for each type of alignment are shown, as well as the absolute value residual. For thickness, we selected the largest residual across the macular location for readability. N, nasal; F, subfoveal; T, temporal.

In Table S5 we see that the significant errors lie within the choroid thickness measurements, and there is no discernible difference in error between the three individuals, regardless of their degree of myopathy. In Fig. S5 we observe the choroidal curvature in all three images (in particular, temporal to the fovea). The degree of this curvature roughly corresponds to the degree of variation of the thickness measurement (green vs. cyan lines) and region of interest definition (shaded green vs additional blue shaded region).

Our results appear to generally contradict our initial hypothesis that high myopia could affect the choroidal measurements. In fact, the primary cause for large differences between choroid-aligned and image-aligned measurements are the extent of deviation of the choroid axis from the image axis, the size of the choroid, and not the extent of myopathy of the eye. The factors which likely contribute to this curvature are imager experience and patient concentration.

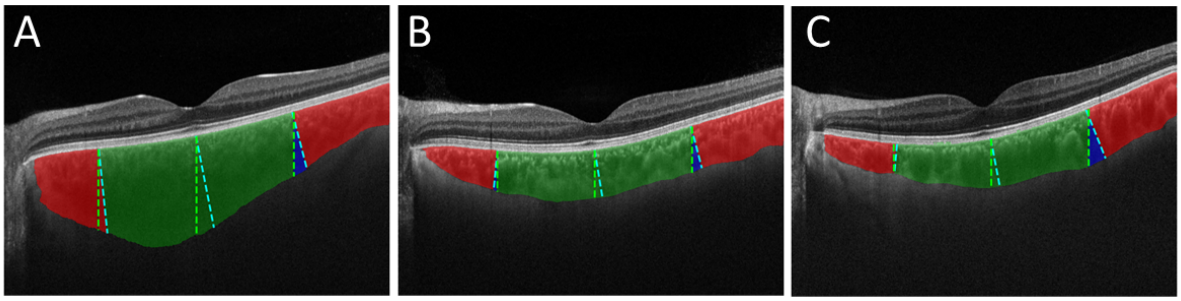


Figure S5 Choroid thickness and area for hypermetropic (A), emmetropic (B) and myopic (C) choroids. Thickness shown as straight lines with green representing image-aligned measurement and cyan as choroid-aligned measurement. Area in shaded green shows overlap between image- and choroid-aligned regions of interest, with blue as regions which were only Choroid-aligned. Shaded red are regions of the choroid not measured within any region of interest.

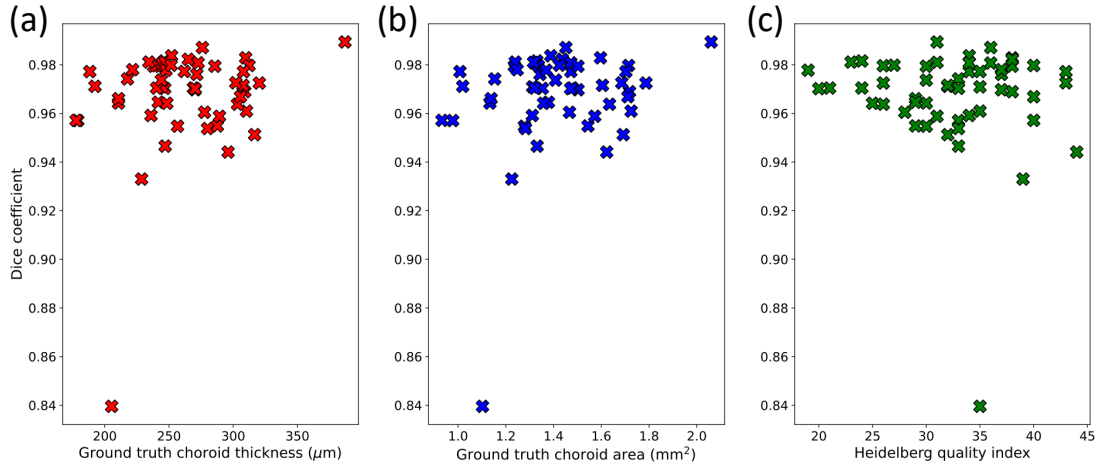


Figure S2: Test set Dice scores plotted against choroid thickness (a), choroid area (b) and Heidelberg-measured quality index (c) in the held-out test set. The outlier Dice score of approximately 0.84 is the dice score between DeepGPET and GPET from figure 4(d).

Figure 6.2: Supplementary Figure S2.

6.3 Conclusion

DeepGPET itself is already a very valuable tool. For instance, it can segment a 61 slice optical coherence tomography volume fully-automatically in a minute or two on a laptop without manual intervention, when previously with GPET would have taken about half an hour with regular human inputs. Choroidalyzer further improves on that substantially by offering vessel related metrics, and identifying the fovea location automatically. In my opinion, the work presented in this chapter represents a valuable contribution to the field that will enable exciting new research.

However, there are many avenues for improvement. First, the model itself could be made more efficient or more robust, or both, through improved training methods, architectures, or expanded datasets. For example, the training procedure and model size could be tuned to make the model even faster downstream use without a loss in performance. Similarly, we could make the model more robust by using more diverse data augmentation, adding an attention block at the lowest internal resolution of the model so it can easily consider the whole image as context, and most importantly by training on more and more diverse data. Second, the data we had available for development and validation was collected for studies relating to systemic health and lacked in ocular pathologies. Investigating whether Choroidalyzer is robust to different types of pathology would be important to understand in what settings it can be employed,

and might highlight shortcomings that could be addressed by re-training with more diverse training data. This includes pathology that majorly affects the choroid itself such as central serous chorioretinopathy as well as other conditions that mainly affect the retina such as age-related macular degeneration, where the retinal layers can become disorganised which could confuse the model. Third, examining the repeatability and robustness to image quality would likewise clarify in what settings it can and cannot be used. Fourth, currently there is no automated quality scoring step involved. If we want to apply Choroidalyzer to large datasets, especially those collected in clinical practice, the choroid might not always be sufficiently well captured, and thus an automated quality assessment step would be essential. Fifth, currently Choroidalyzer requires at least some basic knowledge of Python to use. Ideally, it should be made more accessible, e.g. by providing a graphical user interface.

Chapter 7

Machine learning for efficient automated quality assessment of colour fundus images

7.1 Introduction

Image quality assessment is an important part of retinal image analysis to identify images that are too poor to analyse, both in research and in potential practical applications. Real-time image quality assessment could also be used to provide feedback to camera operators that are not themselves familiar with retinal imaging. For colour fundus imaging, quality assessment is sometimes done manually (Engelmann et al., 2024b) which is labour-intensive and potentially subjective, or with ad-hoc automated methods (Villaplana-Velasco et al., 2023; Zekavat et al., 2022) that might not perform optimally and - if they are not made openly available - hard to reproduce. Some openly available automated methods exist, such as MCFNet (Fu et al., 2019) and AutoMorph (Zhou et al., 2022), both of which use the EyeQ dataset made available by Fu et al. (2019). Both of these methods are a very useful contribution to the field. However, they share two limitations. First, they are somewhat computationally intensive: MCFNet uses a custom model architecture that includes 3 DenseNet121 (Huang et al., 2017) backbones, whereas AutoMorph uses an ensemble of multiple deep learning models. Second, they are not particularly easy to apply. MCFNet relies on code for its custom model architecture as well as a dataloader than inputs three copies of the image into the model, each in a different colour space. AutoMorph is designed as an end-to-end pipeline that is run on a whole dataset at once, so it does not expose functionality to only compute image quality but is set up to also segment the images and compute various retinal traits.

I wanted an efficient, easy-to-use image quality assessment method for my own work, including studying the relationship between fractal dimension and systemic health, investigating the relationship between image quality and repeatability of DART, and to study associations between image quality and subject characteristics. Such a tool might be useful not just for myself but for researchers in the field at large. The work in this chapter is enabled by the work of Fu et al. (2019) who kindly shared the EyeQ dataset they used for training MCFNet. However, the EyeQ dataset classifies images into three categories: good, useable, and bad. Thus, both MCFNet and AutoMorph classify images using these three classes. In practice, I find that unwieldy, as what level of image quality is needed differs depending on the application, and thus a continuous, one-dimensional quality score would often be more convenient. Thus, I set out to develop two versions of an efficient, easy-to-use image quality assessment tool: One providing the same three-way classification as MCFNet and AutoMorph do for direct comparability and for users that prefer this setup, and another providing a continuous quality score that might be more useful in some applications.

7.2 Paper

QuickQual: Lightweight, Convenient Retinal Image Quality Scoring with Off-the-Shelf Pretrained Models

Justin Engelmann[1,2(✉)], Amos Storkey[2], and Miguel O. Bernabeu[1,3]

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, Scotland, UK
justin.engelmann@ed.ac.uk

[2] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK

[3] The Bayes Centre, University of Edinburgh, Edinburgh, Scotland, UK

Abstract. Image quality remains a key problem for both traditional and deep learning (DL)-based approaches to retinal image analysis and identifying poor quality images can be time consuming and subjective. Thus, automated methods for retinal image quality scoring (RIQS) are needed. The current state-of-the-art is MCFNet, composed of three Densenet121 backbones each operating in a different colour space. MCFNet, and the EyeQ dataset released by the same authors, was a huge step forward for RIQS. We present QuickQual, a simple approach to RIQS, consisting of a single "off-the-shelf" ImageNet-pretrained Densenet121 backbone plus a Support Vector Machine (SVM). QuickQual performs very well, setting a new state-of-the-art for EyeQ (Accuracy: 88.50% vs 88.00% for MCFNet; AUC: 0.9687 vs 0.9588). This suggests that RIQS can be solved with generic "perceptual" features learned on natural images, as opposed to requiring DL models trained on large amounts of fundus images. Additionally, we propose a Fixed Prior linearisation scheme, that converts EyeQ from a 3-class classification to a continuous logistic regression task. For this task, we present a second model, QuickQual MEga Minified Estimator (QuickQual-MEME), that consists of only 10 parameters on top of an off-the-shelf Densenet121 and can distinguish between gradable and ungradable images with an accuracy of 89.18% (AUC: 0.9537). Code and model are available on GitHub. QuickQual is so lightweight, that the entire inference code (and even the parameters for QuickQual-MEME) is already contained in this paper.

Keywords: Retinal imaging · Deep learning · Retinal quality scoring

1 Introduction

Retinal colour fundus images are used in ophthalmology for detecting and grading of retinal diseases like diabetic retinopathy, and also capture a detailed picture of the blood vessels, which could be informative about systemic health

A. Storkey and M. O. Bernabeu—Equal supervision.

B. Antony et al. (Eds.): OMIA 2023, LNCS 14096, pp. 32–41, 2023.
https://doi.org/10.1007/978-3-031-44013-7_4

[9,11,12]. However, image quality is a key problem even when data is specifically collected for research purposes. For example, in UK Biobank, recent studies discarded 26% [14] to 43% [11] of the available images due to quality issues and only about 60% of participants were found to have at least one good quality image [8]. However, Retinal Image Quality Scoring (RIQS) can be subjective and even graders with medical backgrounds only have moderate to substantial agreement [7]. Thus, automated RIQS methods are needed to provide objective and reproducible quality scores. Reproducibility is especially as image quality-based exclusions can introduce selection bias by excluding older, male, less-healthy, and non-White subjects more frequently [3]. Even work that develops retinal image improvement [10] or robust retinal image analysis methods [4] depends on reliable quality scores.

Fu et al. [5] introduced an automated RIQS method called MultiColourspaceFusionNetwork (MCFNet) and the EyeQ dataset, a re-annotation of the publicly available Kaggle Diabetic Retinopathy dataset that provides quality annotations on a 3 class scale (Good, Usable, Reject). This work was a huge step forward for the field of RIQS with both MCFNet and the EyeQ dataset being very important contributions in their own right. The authors made the code, model weights, and data annotations publicly available, enabling others to both use and build on their work. However, MCFNet requires specific colourspace data transformation steps and consists of 3 Densenet121 backbones. Thus, MCFNet requires a specific dataloader and model weights, and is a somewhat large model. Recent work showed that "off-the-shelf" DL models pretrained on ImageNet might be able to capture salient information such as age from retinal fundus images even without fine-tuning [2]. Inspired by that, we set out to investigate whether we can develop a simpler yet effective automated RIQS method that uses such an off-the-shelf model with a classical machine learning classifier.

Our main contributions are:

- **QuickQual**, a simple RIQS method based on an "off-the-shelf" Densenet121 and an SVM, that achieves state-of-the-art on EyeQ while requiring only standard libraries and 14 lines of code;
- **Fixed Prior linearisation**, a simple method for converting EyeQ into a continuous task while retaining information about the Usable class;
- **QuickQual-MEME**, an even simpler version of QuickQual with a linear layer instead of an SVM that produces a continuous quality score. In fact, QuickQual-MEME is so lightweight, that the entire code and model parameters are contained in Fig. 5.

2 Methods

2.1 EyeQ Dataset

We use the EyeQ dataset introduced by [5], which provides quality annotations for a subset of the EyePacs Diabetic Retinopathy dataset on Kaggle, with three classes: Good, Usable, Bad. We preprocess the images by removing black areas

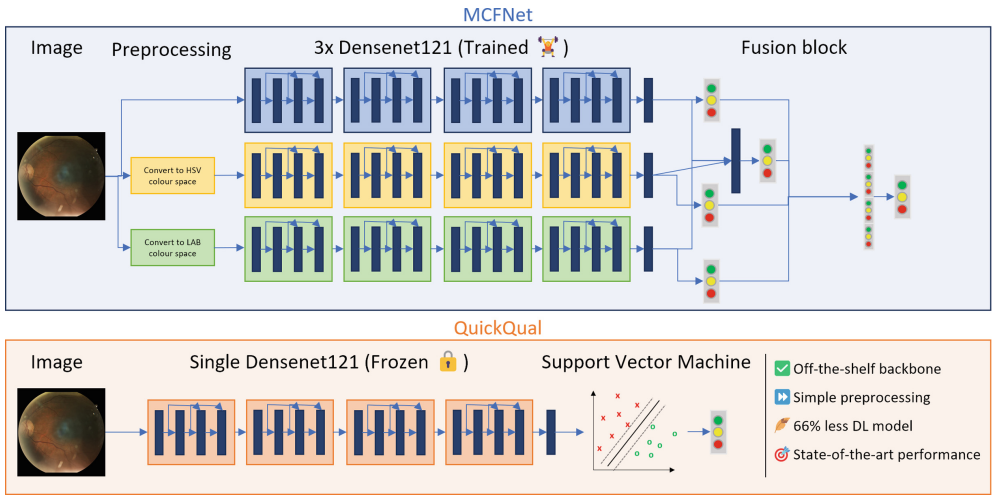


Fig. 1. Comparison between MCFNet (top) and QuickQual (bottom). QuickQual-MEME uses a linear layer instead of SVM.

and then padding the images to square in case they would be non-square otherwise.

2.2 QuickQual

With QuickQual, we aim to develop a method that is quick and convenient to use. By that, we do not merely mean processing speed but also ease of implementation. Our goal is that with less than 20 lines of code and only standard Python libraries, a researcher could apply this method to their own images to obtain quality scores. Thus, we avoid complex preprocessing schemes and non-standard DL architecture code. We use a pretrained DL model from a standard Python DL library and instead of fine-tuning this on the EyeQ dataset, we simply keep it fixed and learn a Support Vector Machine (SVM) on top (Fig. 1).

To enable an easier comparison with MCFNet, we also use Densenet121 [6] as our DL model, but with pre-trained ImageNet [1] weights from the pytorch image models (timm) [13] library. We use a SVM from scikit-learn with standard parameters, except setting "probability=True" to obtain probability scores from the SVM. To obtain discrete class labels, we take the class with the highest probability. We process images at a resolution of 512×512 and simply normalise all channels with mean and standard deviation parameters of 0.5.

2.3 RIQS Beyond 3-Way Classification: Fixed Prior Linearisation

In practice, individual probabilities for three separate classes can be inconvenient to use. Thus, previous work [15] focused on the binary task Gradable (Good or Usable) vs Bad (Reject) instead. This produces a single, continuous score where a simple cut-off for excluding images can be selected. However, this approach treats Good images exactly the same as Usable ones, losing the information that Usable images are at least slightly poorer quality. To remedy this, we propose

a simple linearisation scheme with a fixed prior, i.e. that Usable images are in-between Good and Bad images in terms of quality. During model fitting, we set the optimal output p(Bad) that minimises the loss function to be 0 for Good images, 1 for Bad images, and our fixed prior p for Usable images. In present work, we simply set $p = 0.5$ and thus ask our model to map Usable images in-between Good and Bad ones, thus retaining the information in the labels. This should produce a smooth and desirable quality score but might reduce accuracy for the binary task.

2.4 QuickQual MEga Minified Estimator (QuickQual-MEME)

QuickQual-MEME is an even more lightweight, easy-to-use RIQS model consisting of a pretrained Densenet121 and only 10 parameters for a linear layer. QuickQual-MEME only needs standard python libraries and 15 lines of code. Instead of a SVM, QuickQual-MEME uses a Logistic Regression (Logit) with 10 parameters (9 weights, 1 bias) as classifier. To find these parameters, we proceeded as follows: First, we fit a Logit on the whole EyeQ training set with an L1 penalty ("Lasso") with the default regularisation $C = 1$ and the SAGA optimiser. We then examined the histogram of absolute coefficient magnitudes and chose a cut-off of 0.2 to select 288 of the 1,024 Densenet121 variables. Next, we did forward step-wise features selection using 2-fold crossvalidation on the training set and the AUC as criterion to select the 9 most useful features. Finally, we rounded the parameters to two decimal places so they are easier to report and copy, which led to an insubstantial change in accuracy.

2.5 Evaluation

For the standard EyeQ 3-way classification task, we use standard metrics like Accuracy, F1 score, area under the receiver operating characteristic curve (AUC), logistic loss also known as cross-entropy (LogLoss), cohen's unweighted Kappa (Kappa) and quadratic weighted Kappa (QuadKappa). AUC is a ranking metric that evaluates the model across all possible decision thresholds, whereas LogLoss provides a measure of calibration. Kappa captures how well the model agress with the labels compared to random chance, and QuadKappa penalises errors by more than one class much more, i.e. confusing Good with Bad is worse than confusing Good with Usable. For the binary Gradable vs. Ungradable, we use the same metrics except for Kappa/QuadKappa, which are only suitable for multi-class problems. We calculate all metrics using scikit-learn and use the predicted probabilities for MCFNet provided by the authors to ensure a fair and accurate comparison.[1]

[1] Note that for MCFNet, the original accuracy scores provided were not entirely accurate due to a bug in the evaluation code. See the note here on the Github for MCFNet: https://github.com/HzFu/EyeQ#-reference "*Note: The corrected accuracy score of MCF-Net is 0.8800.*" We thank the authors of MCFNet for their exceptional transparency in sharing not just code, model weights and data, but also their model's test set predictions.

Table 1. Performance for MCFNet and QuickQual on the test set of EyeQ (n=16,249). Note: All metrics are calculated from per-sample predictions using identical code to ensure an accurate comparison. See Sect. 2.5 and footnote 1.

Model	Accuracy	AUC	F1	LogLoss	Kappa	QuadKappa	Filesize
MCFNet [5]	0.8800	0.9588	0.8606	0.3632	0.8017	0.8955	112 MB
QuickQual (ours)	**0.8863**	**0.9687**	**0.8675**	**0.3049**	**0.8107**	**0.9019**	**31 + 25 = 56 MB**

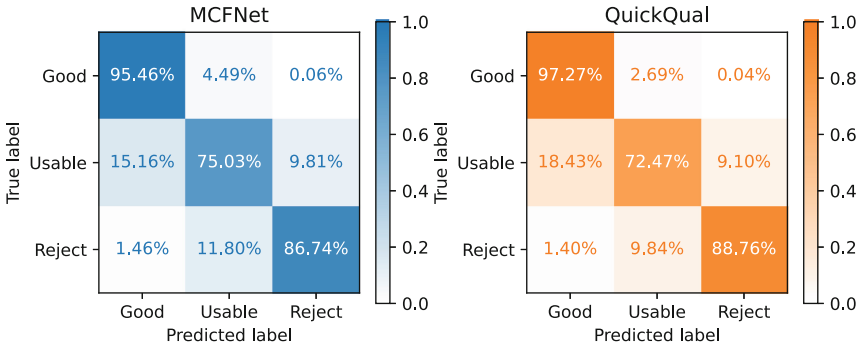


Fig. 2. Confusion matrices for MCFNet and QuickQual, normalised per row.

3 Results

3.1 QuickQual Performance on EyeQ

Table 1 shows the results for QuickQual and MCFNet. QuickQual performs better in every metric. Accuracy, F1 and QuadKappa are slightly better, whereas AUC, LogLoss and Kappa are substantially better. QuadKappa penalises large errors (i.e. confusing Good with Reject) more than Kappa. Thus, QuickQual having a larger improvement in Kappa than in QuadKappa suggests that it is particularly good at distinguishing between the Usable and Good/Reject classes. The confusion matrix (Fig. 2) shows that QuickQual is also better at avoiding large errors (top right and bottom left corners). The only category where QuickQual makes more errors than MCFNet is confusing Usable with Good (middle left). In our opinion, this error is the least concerning type of error - in fact previous work has even combined these two categories [15].

LogLoss is the metric with the largest difference, suggesting that QuickQual is much better calibrated. Fig. 3 shows the distributions of predicted probabilities for both models. Interestingly, MCFNet - unlike QuickQual - never predicts the Usable or Reject classes with large confidence. This might be a by-product of class imbalance and batch training. The QuickQual approach projects the images to small 1,024 dimensional vectors first, which then allows us to fit the SVM to all training images at once.

3.2 QuickQual-MEME Performance on Binary Task

Table 2 shows the results for the binarised task. For comparison, we also evaluate MCFNet and QuickQual on this task, using the predictions for the Reject class,

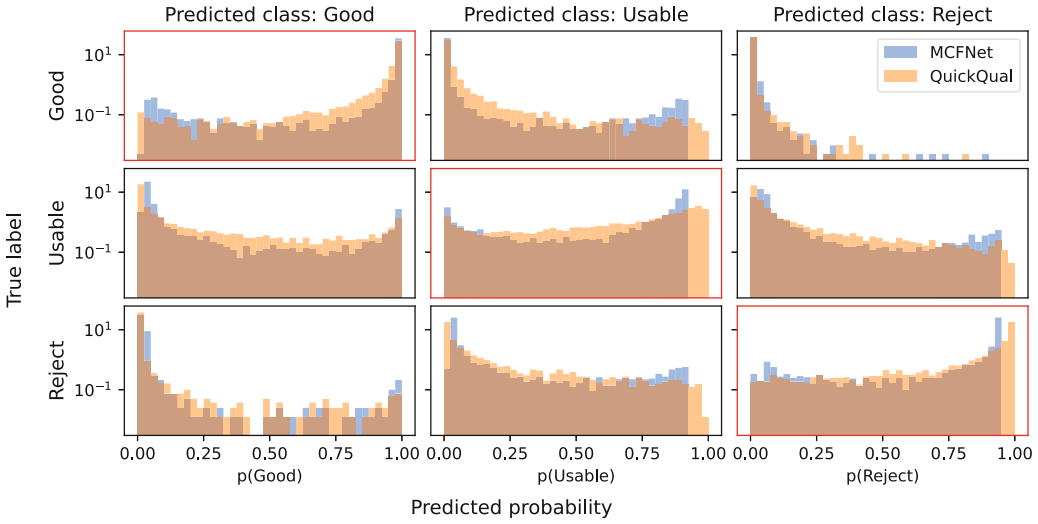


Fig. 3. Distributions of predictions on EyeQ test set for each class, stratified by ground-truth class. Note that y-axis is on a log-scale. This plot is a "soft" version of a confusion matrix. For the diagonal (highlighted red) plots, predictions closer to 1 are better; whereas for the off-diagonal plots, predictions closer to 0 are better. (Color figure online)

Table 2. Performance for binary task Gradable (Good/Usable) vs. Ungradable (Reject).

	Accuracy	AUC	F1	LogLoss
MCFNet [5] (Using p(Reject))	0.9459	0.9819	0.8640	0.1445
QuickQual (Using p(Reject))	**0.9520**	**0.9870**	**0.8799**	**0.1162**
QuickQual-MEME	0.8918	0.9537	0.7602	0.2742
QuickQual-Binary	0.9404	0.9787	0.8505	0.1650

as well as a QuickQual model trained on the binary task. The models trained on the original task perform best, with QuickQual offering slightly better performance in terms of Accuracy and AUC, and a large improvement for F1 and LogLoss over MCFNet. As expected, QuickQual-Binary using the SVM and all 1,024 Densenet121 features outperforms QuickQual-MEME which only uses 9 features.

Interestingly, QuickQual-Binary is outperformed by QuickQual trained on the original task, suggesting that the fixed Prior Linearisation scheme reduces accuracy for Bad vs Good/Usable. However, QuickQual-MEME produces very smooth and desirable quality scores (Fig. 4): The Good and Bad classes have modes on either extremes, while the Usable class is smoothly distributed in-between, with a mode closer to the Good class. This matches the class names: Usable is conceptually closer to Good than to Bad. Images from the Usable class with very low p(Bad) appear to be good quality, while those with high p(Bad) appear poor. Where the distributions of Good and Usable overlap, images are

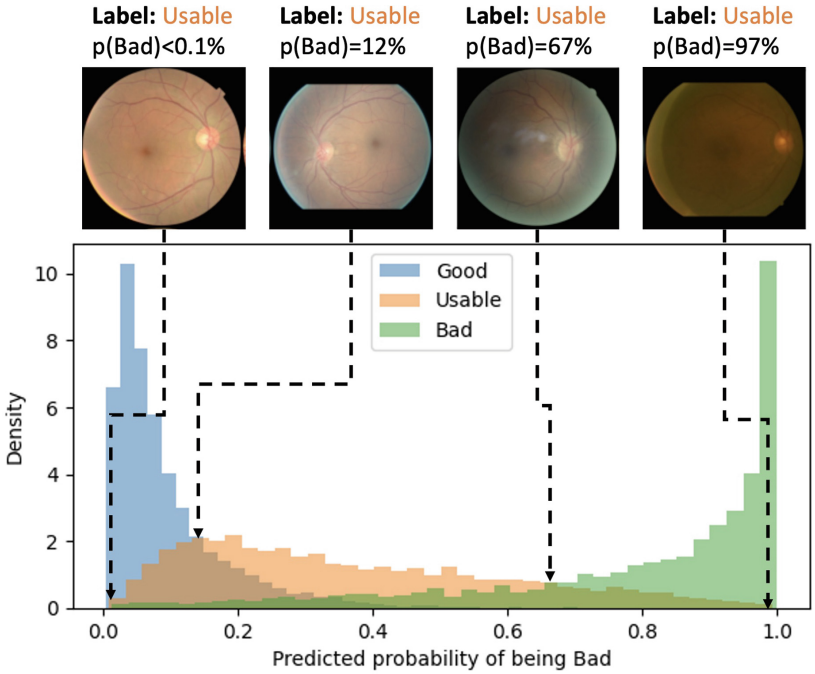


Fig. 4. QuickQual-MEME predicted p(Bad) on the EyeQ test set, stratified by ground-truth class, with example images belonging to the Usable class shown above.

```
import torch
from torchvision.transforms import functional as F
from PIL import Image
import timm
img = Image.open('[DATAFOLDER]/10036_left.jpeg')
model = timm.create_model('densenet121.tv_in1k',
                          pretrained=True, num_classes=0)
model.eval().cuda()
w = torch.tensor([-1411.32, 517.09, 342.41, -707.9,
                  1442.09, -23.25, -541.64, -8.44, 5.44])
b = torch.tensor([5.18])
img = F.to_tensor(F.resize(img, 512))
img = F.normalize(img, [0.5]*3, [0.5]*3).cuda().unsqueeze(0)
with torch.no_grad():
    feats = model(img).squeeze().cpu().reshape(1, -1)
feats = feats[:, [71, 109, 121, 53, 55, 123, 29, 133, 84]]
pred = torch.sigmoid(feats @ w + b)
```

Fig. 5. Entire inference code to run QuickQual-MEME, including the model parameters themselves. The code can be copied from the figure above.

imperfect but still generally good; and where Usable and Bad overlap, they are poor.

Although this evaluation is not comprehensive, this suggests QuickQual-MEME's quality score for the Usable class might align well with actual quality. Giving a very low p(Bad) score to all the Usable images, including the ones that look quite poor, would increase accuracy on the binarised task. However, in our opinion, the current behaviour of QuickQual-MEME appears preferable to that. Thus, accuracy might be an imperfect measure and more fine-grained expert evaluation is needed.

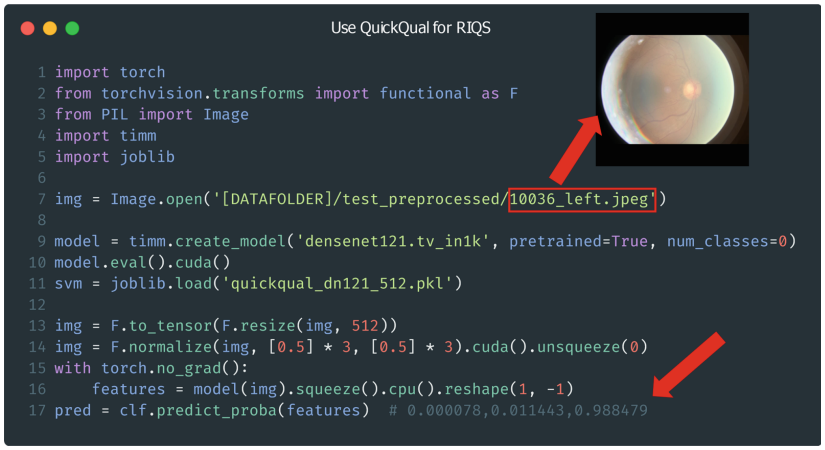


Fig. 6. Entire inference code needed for QuickQual. Arrows highlight the example image which is of poor quality; and the prediction for p(Bad)≈ 99%.

3.3 Convenience and Speed

QuickQual (Fig. 6) and QuickQual-MEME (Fig. 5) need about 15 lines of code to be used together with standard, widely used libraries like PyTorch, scikit-learn and timm. QuickQual-MEME only need 10 parameters to be used, QuickQual needs a 25MB pretrained scikit-learn SVM. This means that QuickQual is very easy to implement and thus very convenient to use for researchers.

Inference times for a single images were measured across 1,000 repetitions, with times reported being mean and standard deviation. QuickQual processed the image in 16.6 ms ± 602 µs on a GPU and 79.5 ms ± 2.45 ms on a CPU. QuickQual-MEME took 14.5 ms ± 536 µs on a GPU and 79.3 ms ± 1.88 ms on a CPU. These times suggest that the SVM only adds minimal overhead compared to a linear model when the Densenet121 is GPU-accelerated and no noticable overhead when no GPU is used. Note that batched inference for multiple images in parallel will likely be even faster per image, but even when processing images one-by-one, 767 images could be processed per minute on a CPU. A time of less than a tenth of a second on a CPU also means that QuickQual could conceivably be deployed in practice to assess images in real time as they are taken.

4 Discussion

We presented QuickQual, which achieves state-of-the-art on EyeQ with only 14 lines of inference code, and QuickQual-MEME which produces a single continuous quality score and fits in Fig. 5. We hope that these will be an easy-to-use, convenient method for other researchers in the field.

We also introduced a Fixed Prior linearisation scheme that better preserves information about the Usable class. While quantitatively this reduced accuracy, limited qualitative evaluation suggests that it might produce a smooth, desirable quality score.

In the future, we plan to evaluate this in more detail by having experts rank images in terms of quality and examining the correlation with QuickQual-MEME's quality score. We also plan evaluate other pretrained DL models to see whether a similarly performant yet more light-weight model could be found that enables even faster computation of quality scores. Additionally, even higher performance might be achieved by training DL models with state-of-the-art architectures for this task. Finally, we plan to externally validate QuickQual and QuickQual-MEME on images from UK Biobank.

Acknowledgements. We thank our friends and colleagues for their help and support. J.E. and this work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
2. Engelmann, J., Storkey, A., Bernabeu, M.O.: Deep learning (dl) identifies age as key axis of perceptual variation in fundus images-without training on fundus images. Investigat. Ophthalmol. Vis. Sci. **64**(9), PB004 (2023)
3. Engelmann, J., Storkey, A., LLinares, M.B.: Exclusion of poor quality fundus images biases health research linking retinal traits and systemic health. Investigat. Ophthalmol. Vis. Sci. **64**(8), 2922 (2023)
4. Engelmann, J., Villaplana-Velasco, A., Storkey, A., Bernabeu, M.O.: Robust and efficient computation of retinal fractal dimension through deep approximation. In: International Workshop on Ophthalmic Medical Image Analysis, pp. 84–93. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16525-2_9
5. Fu, H., et al.: Evaluation of retinal image quality assessment networks in different color-spaces. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 48–56. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_6
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Laurik-Feuerstein, K.L., Sapahia, R., Cabrera DeBuc, D., Somfai, G.M.: The assessment of fundus image quality labeling reliability among graders with different backgrounds. PLoS One **17**(7), e0271156 (2022)
8. MacGillivray, T.J., et al.: Suitability of UK biobank retinal images for automatic analysis of morphometric properties of the vasculature. PLoS One **10**(5), e0127914 (2015)
9. MacGillivray, T., Trucco, E., Cameron, J., Dhillon, B., Houston, J., Van Beek, E.: Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. Br. J. Radiol. **87**(1040), 20130832 (2014)
10. Shen, Z., Fu, H., Shen, J., Shao, L.: Modeling and enhancing low-quality retinal fundus images. IEEE Trans. Med. Imag. **40**(3), 996–1006 (2020)

11. Velasco, A.V., et al.: Decreased retinal vascular complexity is an early biomarker of mi supported by a shared genetic control. medRxiv (2021)
12. Wagner, S.K., et al.: Insights into systemic disease through retinal imaging-based oculomics. Transl. Vis. Sci. Technol. **9**(2), 6 (2020)
13. Wightman, R.: PyTorch Image Models (2019). https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861
14. Zekavat, S.M., et al.: Deep learning of the retina enables phenome-and genome-wide analyses of the microvasculature. Circulation **145**(2), 134–150 (2022)
15. Zhou, Y., et al.: Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. Transl. Vis. Sci. Technol. **11**(7), 12 (2022)

7.3 Conclusion

Despite being very simple, QuickQual obtains state-of-the-art performance on the commonly used EyeQ dataset. However, to me the most important contributions are the ease of using it and the one-dimensional score of the mega minified variant, which is more convenient and useful for many analyses than the three-way classification into good, useable, and bad. QuickQual can be very easy implemented and is efficient to run, which allow for hassle-free assessment of quality for colour fundus imaging datasets. In the work presented in Chapter 3, quality was manually annotated which is time consuming and subjective, whereas QuickQual is automatic and openly available, which could enable more consistent and comparable quality scores across projects. In the future, I intend to use QuickQual to look at the relationship between image quality and the useability of retinal image analysis tool, as in Chapter 4, as well as to investigate factors that drive image quality.

QuickQual could be improved in a number of ways. The EyeQ dataset uses the EyePACS diabetic retinopathy dataset on the Kaggle website, so it primarily contains images of healthy eyes or of eyes with diabetic retinopathy. Future work should validate QuickQual on images with other retinal pathologies. Depending on how well it is found to generalise, it might be useful to retrain it on more diverse datasets, although the EyePACS dataset already seems to contain a variety of different cameras and anecdotally in my own work and that of colleagues, thus far QuickQual has generalised well to different datasets from different populations. A second avenue for improvement would be to develop a more detailed taxonomy for image quality. An aspect of an image is only a quality issue in relation to a specific application. For example, if screening for glaucoma, the optic disc is the primary structure of interest, whereas when screening for age-related macular degeneration the macula is most important. Thus, an image could at the same time be perfectly useable for one task yet unuseable for another. In practice, a one-dimensional quality score is a reasonable approximation, but these edge cases do exist. Thus, a more fine-grained, multi-dimensional quality score could provide a more nuanced way to assess images and to decide which images can and cannot be used for a given application. Finally, a light-weight image quality assessment tool that could be run in real-time could improve image quality during capture, by providing the camera operator with information about the quality of an image they have just taken, as well as identifying what sorts of quality issues appear to have affected the image. The latter could then be used to give recommendations for remedying the issue, for instance if the image was too dark, the flash setting might need adjustment.

Such a tool would be especially useful in screening settings where the camera operator might not be particularly trained or experienced in retinal image capture, e.g. a nurse-technician not focused on ophthalmology. I would like to develop such a system in the future.

Chapter 8

Conclusion

8.1 Summary & reflection

I set out to apply machine learning to retinal image analysis where it might be useful, along three core themes: retinal disease detection, development of retinal image analysis tools, and validation and application of these tools. Each of these themes is exemplified by at least one piece of work contained within this thesis. I developed a disease detection model for ultra-widefield imaging (Chapter 2), three tools for retinal image analysis - computation of retinal fractal dimension from colour fundus images (Chapter 3), analysis of the choroid in optical coherence tomography images (Chapter 6), and quality assessment of colour fundus images (Chapter 7) -, and further validated (Chapter 5) and applied (Chapter 4) one of these tools. Each of these pieces of work has several limitations as discussed in the corresponding chapter.

For the theme of retinal disease detection, the results were promising indicating that the methodology used is effective and that the problem is tractable. Yet without robust external validation, the model cannot be recommended for use in clinical practice. This limitation stemmed from the lack of access to suitable data and I hope to address it in the near future with the next iteration of this work. Of course, after robust retrospective external validation, the model should then be prospectively validated. And even if those results were promising, that would theoretically allow recommending it - in good conscience - for clinical use, but regulatory approval would still be required in practice. This is a challenge I will reflect on in the next section. Thus, while my motivation for this theme was that a retinal disease detection model could be beneficial to clinical practice, this was not achieved as part of this thesis. However, clinical adoption was never thought to be feasible during the PhD period itself, and I think that the work

from Chapter 2 forms the foundation for my future work in this area as well as offering an incremental contribution to the field, demonstrating a more suitable way of framing the problem than what had been considered and showcasing effective, state-of-the-art machine learning methods to tackle it.

For the theme of developing retinal image analysis tools, more tangible impact was achieved, in my own opinion. DART addresses two key limitations of existing tooling, namely lack of robustness and long processing times. The former necessitates exclusion of substantial parts of the available data, which introduces selection bias and reduces sample sizes. The latter makes analysing large scale datasets cumbersome and time-consuming, which slows down research. Yet, despite very promising results, DART follows a novel and thus unproven paradigm compared to existing tools, and requires additional validation. QuickQual provides a more incremental improvement over existing tools compared to DART, fulfilling the same role using the same principle as existing solutions yet is more efficient, easier-to-use and achieves better performance. This more incremental nature likely means that it is easier for practitioners to adopt it, compared to a tool based on a novel paradigm. While “state-of-the-art” results on an existing dataset are highly coveted in machine learning research, the two most important benefits are the ease-of-use and the one-dimensional continuous quality score instead of the “good-useable-bad” classification. Finally, Choroidalyzer provides a comprehensive solution for choroidal analysis and - in my opinion - stands to greatly accelerate research in this area. Many limitations remain, such as the lack of automated quality assessment to allow application to large scale, mixed quality datasets. Nonetheless, it still presents substantial progress and subsequent iterations will address the current limitations. None of these tools have made a tangible impact to clinical practice yet, but I am already aware of a number of researchers from several institutions making use of these tools. The feedback I received to date has been overwhelmingly kind and appreciative. Thus, I think these tools are already making a small, yet tangible impact on research. This will, in turn, will - eventually and indirectly - have a positive impact on clinical practice.

For the theme of application of these tools, much more work is currently ongoing than what is already published and presented in this thesis. Using these tools myself is indeed highlighting both strengths and weaknesses, which inform the next iterations. The application of DART in Chapter 4 provided some encouraging initial results regarding its applicability to primary-care data. However, personally I am particularly proud of the work in Chapter 5. Despite the limitations relating to the datasets and

the fact that the results were surprisingly favourable to a tool that I developed myself, the analysis was done carefully and with thought, and partially addresses a gap in this area. In my opinion, examining the repeatability and robustness of retinal image analysis tools is very important as these tools are the foundation of a vast amount of research. Compared to exciting novel findings with potential real-world applications, such research has a much lower upper bound for potential impact and is of interest to a niche audience due to its technical nature. Yet I believe that it is essential for the health of the whole field of retinal image analysis.

Each of the three themes of this PhD are increasingly indirect in their impact to people's lives, yet important in their own way. Despite the limitations mentioned, my hope is that the work presented in this thesis constitutes a meaningful contribution to the field. Thus, I think it accomplishes what I set out to do during my PhD.

My secondary motivations were to develop and demonstrate proficiency with machine learning methods, to develop a reasonable understanding of retinal imaging, retinal disease and ophthalmic care, and to build collaborations and relationships for a future career. Here, I am quite confident that these was accomplished. I have successfully applied machine learning in a variety of projects and through interactions with colleagues and collaborators learned a lot about the domain I work in. I have presented my work at ophthalmology conferences, visited five hospitals on three different continents, and have collaborators across the world.

8.2 Outlook & future work

Each proper chapter itself already contains a discussion of weaknesses and future work relating to the work presented therein, which I will not repeat here. Instead, I want to focus on some broader themes that I think will be important to tackle in the coming years and decades.

First, tooling for retinal image analysis needs to become more accessible to be widely adopted. By virtue of being openly available and relatively easy to install and use for people that are familiar with the Python programming language, the tools I have developed are already quite accessible compared to the median tool in the field. However, many researchers in ophthalmology and vision research are understandably not familiar with Python. In the future, I would like to develop graphical user interfaces and simple installers for common operating systems to support such potential users.

My impression is that such work is not particularly efficient from an academic career progression perspective but I believe that it would be of great benefit to the field, by allowing researchers to use advanced and effective tools when they otherwise would have relied on time-consuming and subjective manual annotations, ad-hoc and hard to reproduce image processing pipelines, or not been able to do a specific analysis at all.

As explained in Chapter 1, automated disease detection has great potential and is - from a technical perspective - a very tractable problem. However, bringing a model into clinical practice is not merely a technical problem, indeed it is not even a primarily technical problem. There are important questions about how such tools would integrate into existing or newly developed clinical workflows, or what the needs and wants from the clinical side are. But perhaps some of the most important unsolved questions relate to who pays for these models and what business models can sustain their development, implementation and maintenance. Getting regulatory approval for a medical device, which such models would be classified as, is a slow, cumbersome, and expensive process. Once approved, these models need to be adopted in a way that generates sufficient revenue to justify the investment to get to that stage and to sustain the continuous development needed to maintain it, as in practice software needs to be continuously adapted lest it “decays”. One key career goal of mine is to see a system I helped develop implemented in clinical practice prior to my retirement. Given the challenges outlined here, this might prove to be an ambitious goal, despite me having no intention to retire early, health permitting.

Another challenge is robustness and generalisability. Datasets used for research are often carefully curated and tools are evaluated in idealised settings, where poor quality images and borderline or ambiguous cases might have been removed. This leads to very high performance estimates that then do not materialise in practice. Even if the temptation to curate the dataset in a way that makes it easier yet less realistic is resisted, it is hard to assemble diverse datasets from a variety of patient populations and healthcare settings. Despite my opinion that retinal disease detection is generally not very challenging from a technical perspective, it is still not trivial. Anecdotally, I recently visited an experimental screening programme that uses a commercially available machine learning algorithm for disease detection and had the opportunity to get my own eyes assessed by it. The images of one eye were judged to be too poor in quality to assess by the model, despite the nurse-technician’s best effort when capturing them. My other eye was assessed to have age-related macular degeneration, which - according to an ophthalmologist who interpreted the images and examined my

eyes with an ophthalmoscope - appears to be a false positive. While a mere anecdote, it illustrates that there is still work to do to develop robust systems that work in a variety of real-world settings. Robustness of course matters for all types of retinal image analysis tools, not just disease detection. For tools that compute retinal traits, increased robustness means better signal to noise ratio, and thus better statistical power.

Related to robustness and generalisability, fairness is an important issue for retinal image analysis. For tools adopted in clinical practice, we need to make sure that they work well for everyone. The same holds true for tools used in research, if we want our research to be representative and benefit everyone. Initial analyses I have done using UK Biobank data suggest that people that are older, male, non-White, or in poorer health are more likely to be excluded due to image quality. Such exclusions are very common in retinal image analysis but might introduce selection bias. It is also plausible that retinal image analysis work less well for poorer quality images, and thus an association between quality and protected attributes might imply that there is a lack of fairness, even when we do not exclude images. In the future, I would like to investigate this in more detail and try to understand the causal mechanisms behind these associations with image quality.

Finally, using retinal images to better assess the risk of systemic health conditions, or “oculomics” is a field with great potential and excitement. Yet, this excitement is currently primarily driven by finding statistically significant associations between retinal traits and prevalent or incident disease. One inconvenient and perhaps underappreciated truth is that a variable can have a highly significant association (i.e. a p-value that is many orders of magnitude smaller than 0.05) even when adjusting for basic risk factors (e.g. age, sex, smoking status, blood pressure, etc.), yet provide no meaningful increase in predictive performance over these variables. More robust tools for computing retinal traits might increase their signal-to-noise ratio and thus information content, which could be a steps towards realising the potential of oculomics.

Bibliography

Agrawal, R., Ding, J., Sen, P., Rousselot, A., Chan, A., Nivison-Smith, L., Wei, X., Mahajan, S., Kim, R., Mishra, C., et al. (2020). Exploring choroidal angioarchitecture in health and disease using choroidal vascularity index. *Progress in Retinal and Eye Research*, 77.

Balmforth, C., van Bragt, J. J., Ruijs, T., Cameron, J. R., Kimmitt, R., Moorhouse, R., Czopek, A., Hu, M. K., Gallacher, P. J., Dear, J. W., Borooah, S., MacIntyre, I. M., Pearson, T. M., Willox, L., Talwar, D., Tafflet, M., Roubeix, C., Sennlaub, F., Chandran, S., Dhillon, B., Webb, D. J., and Dhaun, N. (2016). Chorioretinal thinning in chronic kidney disease links to inflammation and endothelial dysfunction. *JCI Insight*, 1(20).

Brown, G. C. (1999). Vision and quality-of-life. *Transactions of the American Ophthalmological Society*, 97:473–511.

Burke, J., Engelmann, J., Hamid, C., Moukaddem, D., Pugh, D., Dhaun, N., Storkey, A., Strang, N., King, S., MacGillivray, T., Bernabeu, M. O., and MacCormick, I. J. C. (2024). Domain-specific augmentations with resolution agnostic self-attention mechanism improves choroid segmentation in optical coherence tomography images. *arXiv preprint arXiv:2405.14453*.

Burke, J., Engelmann, J., Hamid, C., Reid-Schachter, M., Pearson, T., Pugh, D., Dhaun, N., Storkey, A., King, S., MacGillivray, T. J., Bernabeu, M. O., and MacCormick, I. J. C. (2023a). An Open-Source Deep Learning Algorithm for Efficient and Fully Automatic Analysis of the Choroid in Optical Coherence Tomography. *Translational Vision Science & Technology*, 12(11):27–27.

Burke, J. and King, S. (2021). Edge tracing using gaussian process regression. *IEEE Transactions on Image Processing*, 31:138–148.

Burke, J., Pugh, D., Farrah, T., Hamid, C., Godden, E., MacGillivray, T., Dhaun, N., Baillie, K., King, S., and MacCormick, I. J. C. (2023b). Evaluation of an automated choroid segmentation algorithm in a longitudinal kidney donor and recipient cohort.

Cheung, C. Y., Thomas, G. N., Tay, W., Ikram, M. K., Hsu, W., Lee, M. L., Lau, Q. P., and Wong, T. Y. (2012). Retinal vascular fractal dimension and its relationship with cardiovascular and ocular risk factors. *American journal of ophthalmology*, 154(4):663–674. e1. ISBN: 0002-9394 Publisher: Elsevier.

Chua, S. Y. L., Thomas, D., Allen, N., Lotery, A., Desai, P., Patel, P., Muthy, Z., Sudlow, C., Peto, T., Khaw, P. T., and Foster, P. J. (2019). Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open*, 9(2):e025077. Publisher: British Medical Journal Publishing Group Section: Epidemiology.

Eghtedar, R. A., Esmaeili, M., Peyman, A., Akhlaghi, M., and Rasta, S. H. (2022). An update on choroidal layer segmentation methods in optical coherence tomography images: a review. *Journal of Biomedical Physics & Engineering*, 12(1):1.

Engelmann, J. and Bernabeu, M. O. (2024). Training a high-performance retinal foundation model with half-the-data and 400 times less compute. *arXiv preprint arXiv:2405.00117*.

Engelmann, J., Burke, J., Hamid, C., Reid-Schachter, M., Pugh, D., Dhaun, N., Moukaddem, D., Gray, L., Strang, N., McGraw, P., Storkey, A., Steptoe, P. J., King, S., MacGillivray, T., Bernabeu, M. O., and MacCormick, I. J. C. (2024a). Choroidalyzer: An Open-Source, End-to-End Pipeline for Choroidal Analysis in Optical Coherence Tomography. *Investigative Ophthalmology & Visual Science*, 65(6):6–6.

Engelmann, J., Kearney, S., McTrusty, A., McKinlay, G., Bernabeu, M. O., and Strang, N. (2024b). Retinal fractal dimension is a potential biomarker for systemic health—evidence from a mixed-age, primary-care population. *Translational Vision Science & Technology*, 13(4):19–19.

Engelmann, J., McTrusty, A. D., MacCormick, I. J. C., Pead, E., Storkey, A., and Bernabeu, M. O. (2022a). Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. *Nature Machine Intelligence*, 4(12):1143–1154. Publisher: Nature Publishing Group.

Engelmann, J., Moukaddem, D., Gago, L., Strang, N., and Bernabeu, M. O. (2024c). Applicability of Oculomics for Individual Risk Prediction: Repeatability and Robustness of Retinal Fractal Dimension Using DART and AutoMorph. *Investigative Ophthalmology & Visual Science*, 65(6).

Engelmann, J., Storkey, A., and Bernabeu, M. O. (2021). Global explainability in aligned image modalities. *arXiv preprint arXiv:2112.09591*.

Engelmann, J., Storkey, A., and Bernabeu, M. O. (2023a). QuickQual: Lightweight, Convenient Retinal Image Quality Scoring with Off-the-Shelf Pretrained Models. In Antony, B., Chen, H., Fang, H., Fu, H., Lee, C. S., and Zheng, Y., editors, *Ophthalmic Medical Image Analysis*, Lecture Notes in Computer Science, pages 32–41, Cham. Springer Nature Switzerland.

Engelmann, J., Storkey, A., and LLinares, M. B. (2023b). Exclusion of poor quality fundus images biases health research linking retinal traits and systemic health. *Investigative Ophthalmology & Visual Science*, 64(8):2922–2922. ISBN: 1552-5783 Publisher: The Association for Research in Vision and Ophthalmology.

Engelmann, J., Villaplana-Velasco, A., Storkey, A., and Bernabeu, M. O. (2022b). Robust and efficient computation of retinal fractal dimension through deep approximation. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 84–93. Springer.

Enoch, J., McDonald, L., Jones, L., Jones, P. R., and Crabb, D. P. (2019). Evaluating whether sight is the most valued sense. *JAMA Ophthalmology*, 137(11):1317–1320.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology*, 186(9):1026–1034. ISBN: 0002-9262 Publisher: Oxford University Press.

Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., and Shao, L. (2019). Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 48–56, Cham. Springer International Publishing.

Huang, F., Dashtbozorg, B., Zhang, J., Bekkers, E., Abbasi-Sureshjani, S., Berendschot, T. T., and ter Haar Romeny, B. M. (2016). Reliability of using retinal vascular fractal dimension as a biomarker in the diabetic retinopathy detection. *Journal of Ophthalmology*, 2016.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, X., Kong, X., Shen, Z., Ouyang, J., Li, Y., Jin, K., and Ye, J. (2023). GRAPE: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. *Scientific Data*, 10(1):520. Number: 1 Publisher: Nature Publishing Group.

Ify Mordi and Emanuele Trucco (2022). The eyes as a window to the heart: looking beyond the horizon. *British Journal of Ophthalmology*, 106(12):1627.

Iovino, C., Pellegrini, M., Bernabei, F., Borrelli, E., Sacconi, R., Govetto, A., Vagge, A., Di Zazzo, A., Forlini, M., Finocchio, L., et al. (2020). Choroidal vascularity index: an in-depth analysis of this novel optical coherence tomography parameter. *Journal of clinical medicine*, 9(2):595.

Kugelman, J., Alonso-Caneiro, D., Read, S. A., Hamwood, J., Vincent, S. J., Chen, F. K., and Collins, M. J. (2019). Automatic choroidal segmentation in oct images using supervised deep learning methods. *Scientific Reports*, 9(1).

Kundu, A., Ma, J. P., Robbins, C. B., Pant, P., Gunasan, V., Agrawal, R., Stinnett, S., Scott, B. L., Moore, K. P., Fekrat, S., et al. (2023). Longitudinal analysis of retinal microvascular and choroidal imaging parameters in parkinson's disease compared with controls. *Ophthalmology Science*, page 100393.

Lemmens, S., Devulder, A., Van Keer, K., Bierkens, J., De Boever, P., and Stalmans, I. (2020). Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: assessment of a potential biomarker. *Frontiers in neuroscience*, 14:16.

Liu, X., Bi, L., Xu, Y., Feng, D., Kim, J., and Xu, X. (2019). Robust deep learning method for choroidal vessel segmentation on swept source optical coherence tomography images. *Biomedical Optics Express*, 10(4).

Luben, R., Wagner, S., Struyven, R., Cortina-Borja, M., Petzold, A., Trucco, E., Mookiah, M. R. K., Rahi, J., Denniston, A. K., and Keane, P. A. (2022). Retinal fractal dimension in prevalent dementia: The AlzEye Study. *Investigative Ophthalmology & Visual Science*, 63(7):4440–F0119–4440–F0119. ISBN: 1552-5783 Publisher: The Association for Research in Vision and Ophthalmology.

MacGillivray, T. J., Cameron, J. R., Zhang, Q., El-Medany, A., Mulholland, C., Sheng, Z., Dhillon, B., Doubal, F. N., Foster, P. J., and Trucco, E. (2015). Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. *PLoS One*, 10(5):e0127914. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, CA USA.

Masumoto, H., Tabuchi, H., Adachi, S., Nakakura, S., Ohsugi, H., and Nagasato, D. (2018a). Retinal detachment screening with ensembles of neural network models. In *Asian Conference on Computer Vision*, pages 251–260. Springer.

Masumoto, H., Tabuchi, H., Nakakura, S., Ishitobi, N., Miki, M., and Enno, H. (2018b). Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. *Journal of Glaucoma*, 27(7):647–652.

Masumoto, H., Tabuchi, H., Nakakura, S., Ohsugi, H., Enno, H., Ishitobi, N., Ohsugi, E., and Mitamura, Y. (2019). Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. *PeerJ*, 7:e6900.

Matsuba, S., Tabuchi, H., Ohsugi, H., Enno, H., Ishitobi, N., Masumoto, H., and Kiuchi, Y. (2019). Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. *International Ophthalmology*, 39(6):1269–1275.

Mazzaferri, J., Beaton, L., Hounye, G., Sayah, D. N., and Costantino, S. (2017). Open-source algorithm for automatic choroid segmentation of oct volume reconstructions. *Scientific Reports*, 7(1).

McGrory, S., Taylor, A. M., Pellegrini, E., Ballerini, L., Kirin, M., Doubal, F. N., Wardlaw, J. M., Doney, A. S. F., Dhillon, B., Starr, J. M., Trucco, E., Deary, I. J., and MacGillivray, T. J. (2018). Towards Standardization of Quantitative Retinal Vascular Parameters: Comparison of SIVA and VAMPIRE Measurements in the Lothian Birth Cohort 1936. *Translational Vision Science & Technology*, 7(2):12.

Muller, J., Alonso-Caneiro, D., Read, S. A., Vincent, S. J., and Collins, M. J. (2022). Application of deep learning methods for binarization of the choroid in optical coherence tomography images. *Translational Vision Science & Technology*, 11(2):23–23.

Nagasato, D., Tabuchi, H., Ohsugi, H., Masumoto, H., Enno, H., Ishitobi, N., Sonobe, T., Kameoka, M., Niki, M., Hayashi, K., et al. (2018). Deep neural network-based method for detecting central retinal vein occlusion using ultrawide-field fundus ophthalmoscopy. *Journal of Ophthalmology*, 2018.

Nagasato, D., Tabuchi, H., Ohsugi, H., Masumoto, H., Enno, H., Ishitobi, N., Sonobe, T., Kameoka, M., Niki, M., and Mitamura, Y. (2019). Deep-learning classifier with ultrawide-field fundus ophthalmoscopy for detecting branch retinal vein occlusion. *International Journal of Ophthalmology*, 12(1):94.

Nagasawa, T., Tabuchi, H., Masumoto, H., Enno, H., Niki, M., Ohara, Z., Yoshizumi, Y., Ohsugi, H., and Mitamura, Y. (2019). Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatment-naïve proliferative diabetic retinopathy. *International Ophthalmology*, 39(10):2153–2159.

Nagasawa, T., Tabuchi, H., Masumoto, H., Enno, H., Niki, M., Ohsugi, H., and Mitamura, Y. (2018). Accuracy of deep learning, a machine learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting idiopathic macular holes. *PeerJ*, 6:e5696.

Ohsugi, H., Tabuchi, H., Enno, H., and Ishitobi, N. (2017). Accuracy of deep learning, a machine-learning technology, using ultra–wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment. *Scientific Reports*, 7(1):1–4.

Pezzullo, L., Streatfeild, J., Simkiss, P., and Shickle, D. (2018). The economic impact of sight loss and blindness in the UK adult population. *BMC Health Services Research*, 18(1):63.

Read, S. A., Fuss, J. A., Vincent, S. J., Collins, M. J., and Alonso-Caneiro, D. (2019). Choroidal changes in human myopia: insights from optical coherence tomography imaging. *Clinical and Experimental Optometry*, 102(3):270–285.

Robbins, C. B., Grewal, D. S., Thompson, A. C., Powers, J. H., Soundararajan, S., Koo, H. Y., Yoon, S. P., Polascik, B. W., Liu, A., Agrawal, R., et al. (2021). Choroidal structural analysis in alzheimer disease, mild cognitive impairment, and cognitively healthy controls. *American Journal of Ophthalmology*, 223:359–367.

Semeraro, F., Morescalchi, F., Russo, A., Gambicorti, E., Pilotto, A., Parmeggiani, F., Bartollino, S., and Costagliola, C. (2019). Central serous chorioretinopathy: pathogenesis and management. *Clinical Ophthalmology*, pages 2341–2352.

Shin, Y. U., Lee, S. E., Kang, M. H., Han, S.-W., Yi, J.-H., and Cho, H. (2019). Evaluation of changes in choroidal thickness and the choroidal vascularity index after hemodialysis in patients with end-stage renal disease by using swept-source optical coherence tomography. *Medicine*, 98(18).

Spaide, R. F., Koizumi, H., and Pozonni, M. C. (2008). Enhanced depth imaging spectral-domain optical coherence tomography. *American journal of ophthalmology*, 146(4):496–500.

Stosic, T. and Stosic, B. D. (2006). Multifractal analysis of human retinal vessels. *IEEE transactions on medical imaging*, 25(8):1101–1107.

Tabuchi, H., Engelmann, J., Maeda, F., Nishikawa, R., Nagasawa, T., Yamauchi, T., Tanabe, M., Akada, M., Kihara, K., Nakae, Y., et al. (2024). Using artificial intelligence to improve human performance: efficient retinal disease detection training with synthetic images. *British Journal of Ophthalmology*.

Tabuchi, H., Masumoto, H., Nakakura, S., Noguchi, A., and Tanabe, H. (2018). Discrimination ability of glaucoma via dcnns models from ultra-wide angle fundus images comparing either full or confined to the optic disc. In *Asian Conference on Computer Vision*, pages 229–234. Springer.

Trucco, E., Ballerini, L., Relan, D., Giachetti, A., MacGillivray, T., Zutis, K., Lupascu, C., Tegolo, D., Pellegrini, E., and Robertson, G. (2013). Novel VAMPIRE algorithms for quantitative analysis of the retinal vasculature. In *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–4. IEEE.

Villaplana-Velasco, A., Pigeyre, M., Engelmann, J., Rawlik, K., Canela-Xandri, O., Tochel, C., Lona-Durazo, F., Mookiah, M. R. K., Doney, A., Parra, E. J., Trucco, E., MacGillivray, T., Rannikmae, K., Tenesa, A., Pairo-Castineira, E., and Bernabeu, M. O. (2023). Fine-mapping of retinal vascular complexity loci identifies Notch regulation as a shared mechanism with myocardial infarction outcomes. *Communications Biology*, 6(1):1–13. Publisher: Nature Publishing Group.

Wagner, S. K., Fu, D. J., Faes, L., Liu, X., Huemer, J., Khalid, H., Ferraz, D., Korot, E., Kelly, C., Balaskas, K., et al. (2020). Insights into systemic disease through retinal imaging-based oculomics. *Translational vision science & technology*, 9(2):6–6.

Wagner, S. K., Hughes, F., Cortina-Borja, M., Pontikos, N., Struyven, R., Liu, X., Montgomery, H., Alexander, D. C., Topol, E., and Petersen, S. E. (2022). AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ open*, 12(3):e058552. ISBN: 2044-6055 Publisher: British Medical Journal Publishing Group.

Wei, X., Sonoda, S., Mishra, C., Khandelwal, N., Kim, R., Sakamoto, T., and Agrawal, R. (2018). Comparison of choroidal vascularity markers on optical coherence tomography using two-image binarization techniques. *Investigative Ophthalmology & Visual Science*, 59(3):1206–1211.

Wong, T. Y., Wang, J. J., Rochtchina, E., Klein, R., and Mitchell, P. (2004). Does refractive error influence the association of blood pressure and retinal vessel diameters? the blue mountains eye study. *American Journal of Ophthalmology*, 137(6):1050–1055.

Xuan, M., Wang, W., Shi, D., Tong, J., Zhu, Z., Jiang, Y., Ge, Z., Zhang, J., Bulloch, G., Peng, G., et al. (2023). A deep learning–based fully automated program for choroidal structure analysis within the region of interest in myopic children. *Translational Vision Science & Technology*, 12(3):22–22.

Yeung, S. C., You, Y., Howe, K. L., and Yan, P. (2020). Choroidal thickness in patients with cardiovascular disease: a review. *Survey of Ophthalmology*, 65(4):473–486.

Zekavat, S. M., Raghu, V. K., Trinder, M., Ye, Y., Koyama, S., Honigberg, M. C., Yu, Z., Pampana, A., Urbut, S., and Haidermota, S. (2022). Deep learning of the retina enables phenome-and genome-wide analyses of the microvasculature. *Circulation*, 145(2):134–150. ISBN: 0009-7322 Publisher: Am Heart Assoc.

Zhou, Y., Wagner, S. K., Chia, M. A., Zhao, A., Woodward-Court, P., Xu, M., Struyven, R., Alexander, D. C., and Keane, P. A. (2022). AutoMorph: Automated Retinal Vascular Morphology Quantification Via a Deep Learning Pipeline. *Translational Vision Science & Technology*, 11(7):12.