UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Towards robust deep learning for medical imaging

Applications in ophthalmology and radiology

de Vente, C.W.

Publication date
2025

Document Version
Final published version

Link to publication

Citation for published version (APA):
de Vente, C. W. (2025). *Towards robust deep learning for medical imaging: Applications in ophthalmology and radiology*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:05 Nov 2026

TOWARDS **ROBUST** DEEP LEARNING FOR MEDICAL IMAGING

APPLICATIONS IN OPHTHALMOLOGY AND RADIOLOGY



COEN DE VENTE

Towards Robust Deep Learning for Medical Imaging

Applications in Ophthalmology and Radiology

Coen de Vente

The work described in this thesis has been carried out within the Quantitative Healthcare Analysis (qurAI) Group of the University of Amsterdam and the Diagnostic Image Analysis Group (DIAG) of the Radboud University Medical Center. The research is supported by Eurostars grant E12712.

Printing: Ridderprint.nl
ISBN: 978-94-6506-705-6

This thesis is licensed under a Creative Commons "Attribution 4.0 International" license, except for Chapter 6, which is licensed under a Creative Commons "Attribution NonCommercial NoDerivatives 4.0 International" license.

Towards Robust Deep Learning for Medical Imaging

Applications in Ophthalmology and Radiology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P. P. C. C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 29 januari 2025, te 16:00 uur

door

Coen William de Vente

geboren te Zaanstad

Promotiecommissie

Promotores:	prof. dr. ir. C. I. Sánchez Gutiérrez	Universiteit van Amsterdam
	prof. dr. B. van Ginneken	Radboud Universiteit Nijmegen
Overige leden:	prof. dr. C. G. M. Snoek	Universiteit van Amsterdam
	prof. dr. C. J. F. Boon	Universiteit van Amsterdam
	prof. dr. J. P. W. Pluim	Technische Universiteit Eindhoven
	prof. dr. C. C. W. Klaver	Erasmus MC
	prof. dr. R. Sznitman	University of Bern
	prof. dr. I. Išgum	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

1 General introduction 1
- 1.1 Deep learning for medical image analysis . . . 2
- 1.2 Robustness and deep learning . . . 3
- 1.3 AI-based COVID-19 analysis in CT . . . 5
- 1.4 AI-based retinal image analysis . . . 6
- 1.5 Thesis outline . . . 10

I Generalizing to in-distribution data 13

2 Automated CO-RADS and chest CT severity scoring 15
- 2.1 Introduction . . . 17
- 2.2 Materials and methods . . . 17
- 2.3 Results . . . 21
- 2.4 Discussion . . . 28
- Appendix 2.A CT protocol . . . 33
- Appendix 2.B Components of the AI system . . . 33
- Appendix 2.C Example output of the CORADS-AI system . . . 35

3 A systematic comparison of COVID-19 grading models 43
- 3.1 Introduction . . . 45
- 3.2 Background . . . 47
- 3.3 Methodology . . . 48
- 3.4 Results . . . 55
- 3.5 Discussion . . . 59
- 3.6 Conclusion and future work . . . 61

II Flagging ungradable and out-of-distribution data 63

4 International challenge on robust glaucoma screening 65
- 4.1 Introduction . . . 67
- 4.2 Datasets . . . 69
- 4.3 Challenge setup . . . 72
- 4.4 Participating methods . . . 74
- 4.5 Results . . . 81
- 4.6 Discussion . . . 85
- 4.7 Conclusions . . . 93

III Mitigating performance drops under data shift 95

5 Reliable classification using Uncertainty-Based Instance eXclusion 97
5.1 Introduction . . . 99
5.2 Related work . . . 101
5.3 Methods . . . 103
5.4 Data . . . 112
5.5 Experimental design . . . 114
5.6 Results . . . 119
5.7 Discussion . . . 122
5.8 Conclusion . . . 130
Appendix 5.A Analysis of most uncertain B-scans . . . 131

6 Generalizable iRORA and cRORA detection models 137
6.1 Introduction . . . 139
6.2 Methods . . . 140
6.3 Results . . . 146
6.4 Discussion . . . 148
Appendix 6.A Additional results . . . 157

7 Standardizing OCT through *en face* conditioned 3D diffusion models 161
7.1 Introduction . . . 163
7.2 Methods . . . 164
7.3 Experiments . . . 168
7.4 Results . . . 169
7.5 Discussion . . . 171
Appendix 7.A Additional dataset and pre-processing details . . . 177
Appendix 7.B Overlapping patches and inpainting . . . 177
Appendix 7.C Additional results . . . 179
Appendix 7.D LPIPS for different anatomical planes . . . 184
Appendix 7.E Generated imaging artifacts . . . 185
Appendix 7.F Ethical approval . . . 185

8 General discussion 187
8.1 Improving robustness . . . 188
8.2 Clinical implications . . . 192
8.3 Democratization and reproducibility through open source science . . . 194
8.4 Concluding remarks . . . 194

Bibliography 196

Summary 219

Nederlandse samenvatting 223

Publications 227

Acknowledgements 235

All we know is still infinitely less than all that still remains unknown.

William Harvey

1

General introduction

1

Human vision likely evolved from chemical processes in single-celled organisms that converted sunlight into energy, to clusters of light-sensitive cells, and finally to the complex retinas and brains that we know today [1]. In his *On the Origin of Species*, Charles Darwin even called the evolution of the eye "absurd in the highest possible degree" due to its extraordinary complexity [2]. Humanity has been striving to replicate this highly sophisticated human vision by teaching machines to interpret digital images, a major challenge in the field of computer vision. Computer vision is often said to have started out as a – rather ambitious – student summer project in 1966 at MIT [3]. Ever since, computer vision and artificial intelligence (AI) have seen major advancements and achievements. An important and popular application of AI has been the interpretation of medical images, such as for the identification of diseases in images of the eye itself. This completes a fascinating cycle: human vision inspires computer vision, which, in turn, aids in curing human vision. However, just as human vision and health have their shortcomings and vulnerabilities in the form of diseases, computer vision models can also fail in unexpected ways.

This thesis focuses on when such failures occur in AI for medical imaging and proposes solutions to mitigate these. The concepts related to these challenges are introduced in Sections 1.1 and 1.2. The initial goal of this thesis was to develop systems for the automated analysis of ophthalmic images to support the diagnosis and management of retinal diseases. However, the onset of the coronavirus disease 2019 (COVID-19) pandemic during the first year of this project led to our participation in a collaboration among research groups from several Dutch hospitals, aimed at developing AI-based solutions for the analysis of COVID-19 in CT. This collaboration not only addressed an urgent clinical need but also resulted in practical, technical, and scientific developments related to deep learning-based medical image analysis. As a result, this thesis investigates the robustness of AI for medical imaging in the context of COVID-19 analysis using computed tomography (CT), introduced in Section 1.3, and retinal imaging, introduced in Section 1.4. Section 1.5 provides an outline of this thesis.

1.1 Deep learning for medical image analysis

Over the last decade, deep learning (DL) has substantially transformed the field of AI. DL is a collection of methods based on artificial neural networks, with major breakthroughs every few years – if not more often [4–12]. Although other forms of machine learning may still be at least non-inferior in certain cases [13], the capability of DL models to effectively learn from vast amounts of information [11, 12, 14] and their consistent top-ranking po-

sitions in machine learning benchmarks* make them very popular model choices. The widespread adoption is further supported by their relative ease of development, facilitated by well-maintained, open-source frameworks [15–17] and the fact that they do not require hand-crafted features.

DL has revolutionized medical image analysis too [18]. It has become the norm for processing images in numerous medical specializations, such as radiology [19], ophthalmology [20], and histopathology [21]. More and more in-the-lab AI successes lead to products approved for clinical use [22]. With this trend of shifting towards applications in the real world, however, we should be wary of how models can fail unexpectedly during deployment.

1.2 Robustness and deep learning

A study by Beede *et al.* illustrates how a DL algorithm that demonstrated promising results in controlled lab settings, encountered unforeseen challenges when implemented in real-world environments [23]. They applied an AI system for screening diabetic retinopathy using retinal images within a real screening setting in Thailand. A major issue was the system's strictness about image gradability, which led to a high rate of images being flagged as ungradable (21% of all images). This had negative effects on the usability of the system and user satisfaction. Nurses interacting with the system provided comments about the DL tool, such as:

> *"It gives guaranteed results, but it has some limitations. Some images are blurry, and I can still read it, but [the system] can't."* [23]

The discrepancy between expected and actual performance in this scenario can be primarily attributed to differences in data acquisition conditions. Specifically, poorer lighting conditions during real-world deployment contrasted with those in laboratory settings, exemplifying the necessity and relevance of researching model robustness.

Many definitions of robustness can be found within the machine learning literature. Research works often do not explicitly specify a definition of robustness when they use the term, or they use a more specific definition of robustness, such as insensitivity to adversarial attacks. To address this variability in definitions, Freiesleben *et al.* provide a general definition of robustness that is broadly applicable across various scientific domains, yet specifically developed for machine learning models:

*https://paperswithcode.com/sota/image-classification-on-imagenet

1

> *"The robustness target is said to be robust to the robustness modifier if relevant interventions in the modifier, as specified by the robustness domain, do not lead to greater changes in the target than specified by the target tolerance."* [24]

This definition describes a causal relationship between the "robustness target" and the "robustness modifier". In the context of machine learning models, the target is often the model performance during deployment and the modifier is often the test data. Robustness targets can also refer to other model properties such as explanations, and robustness modifiers can refer to matters such as the training algorithm or social dynamics [24]. Using this definition as a foundation, we refer to robustness as *a model's stability – in terms of any performance metric of interest – under variations in the test data distribution* in the remainder of this thesis, unless stated otherwise. With this definition in place, we can highlight several important facets of robustness and examine their implications on DL for medical image analysis.

Generalization to distributions close or identical to the training distribution. A main objective of machine learning research is the generalization to distributions identical or similar to the training distribution. Issues regarding this ability can arise when the training and test datasets are not representative of the underlying data distribution [24].

This is particularly relevant in medical imaging, where obtaining enough data from a sufficiently large number of patients can be difficult [25]. To address this, common strategies include regularization approaches such as data augmentation, which involves applying random image transformations such as rotations and contrast changes [24]. Well-tuned training pipelines with such straightforward techniques provide preferable alternatives to complex, problem-specific solutions that may not generalize well across different datasets [26, 27].

Data shifts. Data shift, also known as distribution shift, occurs when the distribution of test or deployment data diverges from that of the training data. This divergence can lead to performance degradation at test time. Various factors including imaging artifacts, noise level variations, or image resolution changes can cause data shifts.

A prominent example of data shift can be observed in workflows for the automated analysis of retinal optical coherence tomography (OCT), where the plurality of protocols and device manufacturers leads to diverse image characteristics, as we will discuss further in Section 1.4.1. A possible mitigation strategy is transfer learning, in which the model continues training on data from the new distribution. However, it is not always known when data has undergone a distribution shift. Moreover, this approach necessitates the collection of new data and corresponding manual labels every time the model is applied to a new environment, which can be costly and unfavorable for the scalability of the system. Ad-

ditionally, there is the risk of catastrophic forgetting, where the model forgets previously acquired knowledge. Therefore, developing DL models that are inherently robust to data shifts, requiring fewer resources and operating seamlessly under varying conditions, is imperative.

Out-of-distribution detection. It is generally considered ideal for models to maintain stable performance under distribution shifts. However, occasionally, this is infeasible due to the shift being too large. In such cases, it can already be sufficient to recognize that the data distribution underwent such a shift, a practice known as out-of-distribution (OOD) detection [28].

Within clinical workflows, the notification of an OOD sample could lead to a corrective action such as the reacquisition of a scan or the reconsideration of the acquisition protocol, instead of the model silently failing. The task of identifying ungradable images, such as those in the diabetic retinopathy screening program discussed at the beginning of this section, could also be formulated as an OOD detection problem.

1.3 AI-based COVID-19 analysis in CT

The global COVID-19 pandemic presented unprecedented challenges to healthcare systems worldwide. Outbreaks overwhelmed hospitals and led to shortages in the availability of viral testing using reverse transcription-polymerase chain reaction (RT-PCR) tests. As RT-PCR assays could take up to multiple days [29] and COVID-19 rapid antigen tests were not yet available in the first stages of the pandemic, chest CT became an additional highly valuable instrument for the triage of COVID 19-suspected patients [30]. CT imaging allows for the diagnosis and severity quantification of COVID-19 through, among other features, the assessment of multifocal ground glass opacities and consolidations in the lungs [31].

Standardized systems, such as COVID-19 Reporting and Data System (CO-RADS) [29], were developed for the interpretation and reporting of chest CT findings in COVID-19 cases, aiming to improve communication among healthcare workers. AI solutions to automatically grade CT scans with these standardized systems could help radiologists to reduce their workload and potentially improve their accuracy. While numerous studies have focused on AI-based COVID-19 detection in CT [31–46], external evaluation, an important practice for evaluating the robustness of AI systems, or adherence to standardized systems for the reference standard were often lacking.

The development of AI solutions for COVID-19 CT analysis not only addressed an immediate clinical need, but we think it also accelerated advancements in the broader field of AI for medical image analysis. The urgency of the pandemic has led to efficient collabora-

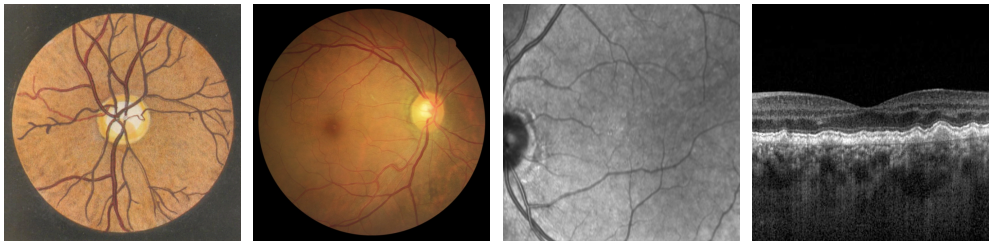
tions between healthcare professionals and engineers. Outcomes included generalizable methods for the automated quantification and grading of diseases and the further development of the online platform *grand-challenge.org*, including its capabilities allowing for the collaborative annotation of 3D medical data, efficient algorithm sharing and evaluation through public challenges. The features of this platform, which were actively being developed during the writing of this thesis, were instrumental in addressing the practical requirements of our research.

1.4 AI-based retinal image analysis

1.4.1 Fundus imaging

The first published image of the interior of the eye is often attributed to Adrien van Trigt, who included hand-drawn images of the retina in his doctoral thesis in 1853 [47–49]. He drew these images based on what he saw using the ophthalmoscope, which had been invented a few years earlier [50]. An example of one of these images is shown in Fig. 1.1a. This is a *fundus* or *en face* image, *i.e.,* a 2D representation of the 3D, semi-transparent tissues of the retina, captured using reflected light [49]. Fundus imaging has undergone numerous developments within the last 170 years, leading to substantially improved images.

Combining ophthalmoscopy and digital photography, we are now able to produce and digitally analyze images called color fundus photographs (CFPs) (see Fig. 1.1b). Many anatomical retinal structures can be made visible with CFPs, such as the macula and the optic disc. The macula is the central region of the retina that contains the fovea, a small pit in the middle of the macula that provides sharp central vision. The optic disc is the location in the retina where the axons of retinal neurons leave the brain, delivering signals



(a) (b) (c) (d)

Figure 1.1: A brief overview of the evolution of retinal imaging. **(a)** The first published "color fundus image", drawn by Adrien van Trigt in 1853, as part of his doctoral thesis [47]. Courtesy of the *Utrecht University Library*. **(b)** A color fundus photograph. **(c)** A scanning laser ophthalmoscopy (SLO) image. **(d)** A B-scan from an optical coherence tomography (OCT) volume.

to the brain. In Section 1.4.3, we discuss two prevalent diseases that specifically impact the macula and the optic disc.

Confocal scanning laser ophthalmoscopy (SLO) is another important retinal imaging technique that produces *en face* images, invented in the early 1980s [51, 52]. It is based on a laser beam that moves over the retina in a 2D grid pattern, allowing for highly exact spatial localization with minimal image obstruction from scattered light. An example of an SLO image is shown in Fig. 1.1c.

1.4.2 Optical coherence tomography

OCT is another retinal imaging modality and has become commonplace in ophthalmic clinics since the start of this millennium [53]. OCT is a fast and non-invasive way of generating 3D cross-sectional images of the retina, often considered the gold standard for diagnosing retinal diseases [54]. It works by estimating the depth of tissue based on the time

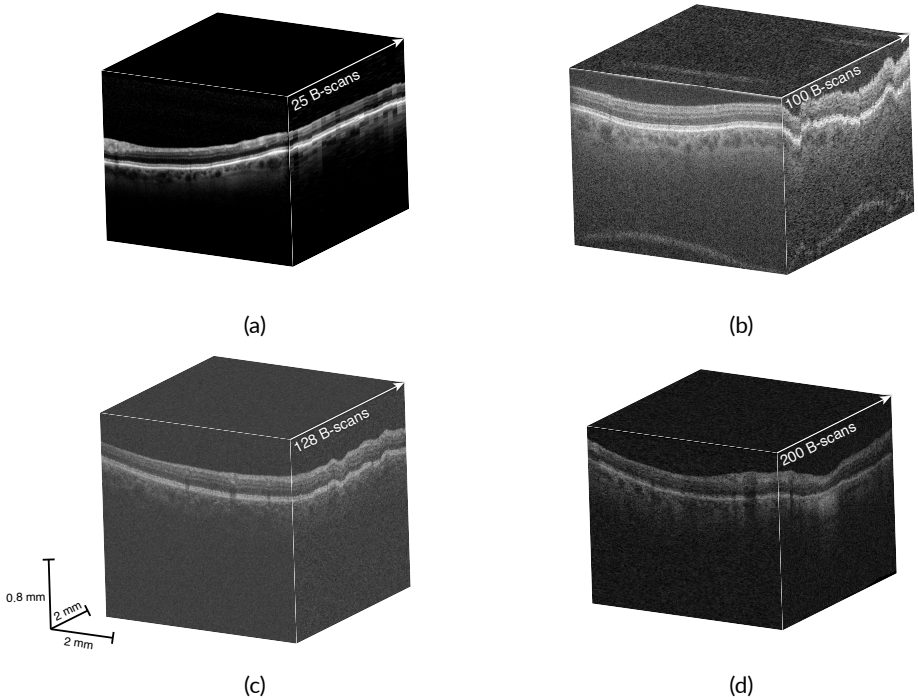


Figure 1.2: Optical coherence tomography (OCT) scans acquired with different imaging systems from **(a)** Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany), **(b)** Bioptigen, Inc (Research Triangle Park, North Carolina, United States), **(c)** Topcon Corp. (Tokyo, Japan), and **(d)** Cirrus (Zeiss, Meditec, Oberkochen, Germany). OCT volumes are displayed as cubes, illustrating the variety of B-scan spacing and signal-to-noise ratio in retinal OCT imaging.

it takes for backscattered light to return to a sensor, through interferometry [49]. A light beam scans the retina in a 2D raster pattern, similar to the acquisition of an SLO image. For each point on this raster, an *A-scan* is produced, which is a 1D representation of the signal across various depths. A set of laterally aligned A-scans forms a 2D image, referred to as a *B-scan* (see Fig. 1.1d). A collection of stacked B-scans is referred to as an OCT volume (or a *C-scan*). Some scanners utilize SLO imaging as a reference for localization, while acquiring the OCT scan [55]. This allows for eye movement corrections during OCT acquisition, but also for easier registration of the OCT volume with the SLO image itself and other *en face* modalities, such as CFPs. OCT imaging allows for distinguishing retinal layers and identifying structural abnormalities of the retina, which is highly valuable when detecting and monitoring diseases.

Since the introduction of OCT, many device manufacturers have entered the ophthalmic OCT market. Swanson *et al.* list fourteen companies that were in this space before 2015 [56]. This has led to a diverse array of scanning protocols and image characteristics across different devices. A primary source of this heterogeneity is the technique of acquiring multiple B-scans at the same location and averaging them, which improves the signal-to-noise ratio. The inconsistent use of B-scan averaging across devices and protocols leads to considerable variations in image noise levels (see Fig. 1.2). Moreover, B-scan averaging substantially increases acquisition time, so OCT volumes acquired using this method generally comprise fewer B-scans.

Additional factors contributing to inter-scanner variability include differences in how imaging artifacts are managed. For example, some devices automatically retake B-scans affected by blinking, while others do not, resulting in certain artifacts that are only present in OCT scans from specific devices. These variations in OCT acquisition protocols and image characteristics present challenges in data standardization, the comparability of results across and within clinical studies, and AI model deployment.

1.4.3 Retinal diseases, their manifestation in imaging, and how AI can play a role

The overall well-being of a population is heavily influenced by eye health [57]. It has been estimated that worldwide 43 million people are blind and 295 million people have moderate to severe vision impairment [58]. On a positive note, however, many forms of blindness can be prevented or their progression can be cost-effectively inhibited [57]. Leading causes of blindness include age-related macular degeneration (AMD) and glaucoma, diseases that both affect the retina. AI can play an important role in the screening, diagnosis, and monitoring of these conditions.

Age-related macular degeneration. AMD is a medical condition affecting central vision,

caused by a progressive deterioration of the macula. AMD has a worldwide prevalence of 8.7% [59] and is the fourth most common cause of blindness [57, 60]. AMD can be categorized into early AMD, intermediate AMD, and late AMD. Clinical grading systems such as the Beckman classification system define these stages [61]. This system, published in 2013 and established using fundus photography, defines these stages based on the presence of (various sizes of) drusen, pigmentary changes, neovascularization, and geographic atrophy (GA). The two types of late (or advanced) AMD are wet AMD and dry AMD. The former is characterized by neovascular lesions (new blood vessels) and the latter by GA (atrophy of the retinal pigment epithelium (RPE)). More recently, the Classification of Atrophy Meeting (CAM) group established more granular definitions for patients with GA, called incomplete RPE and outer retinal atrophy (iRORA) and complete RPE and outer retinal atrophy (cRORA) [62]. The definitions of iRORA and cRORA are based on the degradation of specific retinal layers visible in OCT.

Multiple treatments for wet AMD have been available for many years, such as laser photocoagulation and intravitreal injections of anti-vascular endothelial growth factor (anti-VEGF) [63]. A significant advancement in AMD treatment occurred in 2023, with the United States Food and Drug Administration (FDA) approval of the first drug for dry AMD treatment, marking a major milestone in the prevention of vision loss [64]. Retinal imaging plays an important role in guiding treatment decisions and evaluating treatment response [65]. Traditionally, clinicians have relied on the manual inspection of images to assess disease progression. However, precise quantification of disease-specific features, such as fluid volume in wet AMD patients and GA area in dry AMD patients, particularly in volumetric OCT data, is cumbersome and time-consuming. Consequently, such quantitative analyses are often omitted in clinical practice, with disease progression typically estimated through visual assessment. This approach is inherently more subjective and less accurate compared to pixel-level quantification.

To address these limitations, there has been significant interest in the development of automated segmentation algorithms for AMD-related features [66–69]. These studies have demonstrated that AI-based approaches can perform at least as well compared to human experts in the quantification of AMD-related features, showcasing their potential to enhance the efficiency and effectiveness of AMD management.

Glaucoma. Glaucoma is a collection of conditions that lead to the deterioration of the optic nerve, affecting peripheral vision. It is the third most common cause of blindness [57] and has been estimated to affect 111.8 million people by 2024 [70]. The delayed detection of glaucoma is the main reason for severe visual loss and blindness [71, 72]. While the damage caused by this condition is irreversible, progression can be effectively stopped through

cost-effective treatments that decrease intraocular pressure, such as eye drops or laser therapy [73]. Therefore, screening could play a crucial role in preventing this disease, particularly in developing countries, where only 5 to 10% of glaucoma cases are diagnosed [74]. CFP-based screening programs may be a cost-effective strategy, allowing for a simple and relatively inexpensive way of imaging the optic disc [74]. Robust AI systems for the automated analysis of CFPs could significantly enhance the economic and practical viability of glaucoma screening programs by alleviating the need for human experts, which are especially scarce in developing countries.

1.5 Thesis outline

This thesis focuses on three facets related to the robustness of DL models for medical image analysis, specifically for COVID-19 grading in CT scans and the assessment of AMD and glaucoma with retinal imaging. Part I focuses on the ability of these models to perform consistently across input data distributions close to or identical to the training distribution. Part II considers the identification of ungradable samples through OOD detection. Part III addresses the mitigation of performance drops under data shift, when the deployment data is moderately or locally unaligned with the training data. The following research questions are addressed:

Research question 1: *Can we develop a DL system for medical image classification that performs well on external data closely following the training distribution?*

The generalization to data similar or identical to the training distribution is a rudimentary requirement of most deep learning models. Yet, in medical image analysis in particular, achieving and evaluating this requirement is not always trivial. We hypothesize that we can maintain good performance on external data closely following the training distribution by leveraging various popular DL architectures and methodologies. In Chapter 2, we present CORADS-AI, a DL system for the evaluation of pulmonary COVID-19 in chest CT scans through CO-RADS and CT severity scoring. In Chapter 3, we perform a more in-depth technical evaluation of the CO-RADS classification model, comparing several commonly used architectures and methodologies. The models are trained and internally tested on data from the same distribution, from one medical center. We evaluate with two external test sets from other centers, which exhibit minor differences in data characteristics compared to the training data.

Research question 2: *Can we develop a robust medical image classification pipeline that flags ungradable data through OOD detection?*

The application of DL models to OOD or ungradable data can lead to substantial per-

formance degradation or silent failure in real-world settings. Therefore, identifying when such data is presented is highly important in medical image analysis. By publicly releasing a large-scale dataset with labeled and gradable images, we expect that we can encourage the research community to develop robust methods for the identification of OOD and ungradable data. Therefore, in Chapter 4, we introduce the Artificial Intelligence for RObust Glaucoma Screening (AIROGS) challenge. With a diverse dataset comprising approximately 113,000 CFPs from 60,000 patients across 500 screening sites, we aim to accelerate the development of AI solutions for glaucoma screening that generalize well in a wide variety of deployment settings. By only allowing the use of gradable data to develop their solutions, participants are encouraged to design robust OOD detection methods to identify ungradable images. We evaluate the generalization of the participating methods on three external datasets.

Research question 3: *Does suppressing the influence of image parts associated with high uncertainty improve model robustness to OOD imaging artifacts?*

When individual images in the deployment data are only partially OOD, *e.g.,* due to imaging artifacts, model performance is often still affected for those samples. We hypothesize that we can mitigate this compromised performance by identifing the OOD parts in the image using uncertainty estimation techniques and suppressing their influence on the final model prediction. In Chapter 5, we propose Uncertainty-Based Instance eXclusion (UBIX), an approach to improve the robustness of deep learning classification models to locally OOD data. During inference, UBIX leverages multiple-instance learning and uncertainty estimation to reduce the influence of corrupted image crops on the final classification output. We train on data from a single OCT device, for the task of AMD staging in OCT, and evaluate on external datasets acquired with different devices that included vendor-specific artifacts.

Research question 4: *Can domain adaptation mitigate performance drops due to acquisition device-induced data shifts?*

A common cause of data shift in medical imaging is the use of a different image acquisition device during model deployment than during training. We expect that we can employ domain adaptation (DA) techniques to keep up model performance as a result of such shifts. In DA, we aim to align the source distribution to a target distribution, which, in our case, are the deployment and the training distribution, respectively. In Chapter 6, we explore the use of DA to enhance the robustness of segmentation models for iRORA and cRORA lesions. We utilize a segmentation model trained on data from a single OCT scanner. We employ a Cycle-Consistent Generative Adversarial Network (CycleGAN) to adapt the visual characteristics of images from a new scanner to resemble those of the training

data. This method is externally evaluated using data from MACUSTAR, a large prospective European and multicenter study focusing on intermediate AMD and its progression.

Research question 5: *Can we generate standardized, less anisotropic data from highly anisotropic, volumetric medical images by conditioning 3D diffusion models on complementary image modalities?*

Variability in the isotropy of image data, *i.e.,* the variation in resolutions of the anatomical planes, can cause data shifts in volumetric medical imaging. This can lead to performance degradation in models optimized using imaging datasets with limited resolution variability. We expect that we can leverage recent advancements in generative deep learning to address this variability across and within studies. Specifically, in Chapter 7, we develop a super-resolution 3D diffusion model to reduce the anisotropy of retinal OCT volumes. We propose to condition these models on *en face* imaging, enabling the generative model to make well-informed decisions that adhere to the biological truth. Although the approach could be extended to other *en face* modalities, in our experiments, we utilize SLO imaging as conditioning information, which is often concurrently acquired by the OCT device and used internally as a reference for OCT acquisition.

An overview of the publications on which the thesis chapters are based can be found in the Publications section, at the of this thesis. This overview includes author contributions, related contributed resources (source code, algorithms, and challenges), and other publications to which the author has contributed.

PART

Generalizing to in-distribution data

2

Automated CO-RADS and chest CT severity scoring

Authors: Nikolas Lessmann*, Clara I. Sánchez*, Ludo Beenen, Luuk H. Boulogne, Monique Brink, Erdi Calli, Jean-Paul Charbonnier, Ton Dofferhoff, Wouter M. van Everdingen, Paul K. Gerke, Bram Geurts, Hester A. Gietema, Miriam Groeneveld, Louis van Harten, Nils Hendrix, Ward Hendrix, Henkjan J. Huisman, Ivana Išgum, Colin Jacobs, Ruben Kluge, Michel Kok, Jasenko Krdzalic, Bianca Lassen-Schmidt, Kicky van Leeuwen, James Meakin, Mike Overkamp, Tjalco van Rees Vellinga, Eva M. van Rikxoort, Riccardo Samperna, Cornelia Schaefer-Prokop, Steven Schalekamp, Ernst Th. Scholten, Cheryl Sital, J. Lauran Stöger, Jonas Teuwen, Kiran Vaidhya Venkadesh, Coen de Vente, Marieke Vermaat, Weiyi Xie, Bram de Wilde, Mathias Prokop, Bram van Ginneken (*equal contribution)

Original title: Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence

Published in: Radiology (Volume: 298, Issue: 1, Pages: E18-E28, January 2021)

https://doi.org/10.1148/radiol.2020202439

Abstract

Background: The coronavirus disease 2019 (COVID-19) pandemic has spread across the globe with alarming speed, morbidity, and mortality. Immediate triage of patients with chest infections suspected to be caused by COVID-19 using chest CT may be of assistance when results from definitive viral testing are delayed.

Purpose: To develop and validate an artificial intelligence (AI) system to score the likelihood and extent of pulmonary COVID-19 on chest CT scans using the COVID-19 Reporting and Data System (CO-RADS) and CT severity scoring systems.

Materials and methods: The CO-RADS AI system consists of three deep-learning algorithms that automatically segment the five pulmonary lobes, assign a CO-RADS score for the suspicion of COVID-19, and assign a CT severity score for the degree of parenchymal involvement per lobe. This study retrospectively included patients who underwent a nonenhanced chest CT examination because of clinical suspicion of COVID-19 at two medical centers. The system was trained, validated, and tested with data from one of the centers. Data from the second center served as an external test set. Diagnostic performance and agreement with scores assigned by eight independent observers were measured using receiver operating characteristic analysis, linearly weighted κ values, and classification accuracy.

Results: A total of 105 patients (mean age, 62 years ± 16 [standard deviation]; 61 men) and 262 patients (mean age, 64 years ± 16; 154 men) were evaluated in the internal and external test sets, respectively. The system discriminated between patients with COVID-19 and those without COVID-19, with areas under the receiver operating characteristic curve of 0.95 (95% CI: 0.91, 0.98) and 0.88 (95% CI: 0.84, 0.93), for the internal and external test sets, respectively. Agreement with the eight human observers was moderate to substantial, with mean linearly weighted κ values of 0.60 ± 0.01 for CO-RADS scores and 0.54 ± 0.01 for CT severity scores.

Conclusion: With high diagnostic performance, the CO-RADS AI system correctly identified patients with COVID-19 using chest CT scans and assigned standardized CO-RADS and CT severity scores that demonstrated good agreement with findings from eight independent observers and generalized well to external data.

2.1 Introduction

During the coronavirus disease 2019 (COVID-19) pandemic, chest CT imaging has been found useful in the diagnosis and follow-up of patients with COVID-19 [30]. Standardized CT scoring systems, such as the COVID-19 Reporting and Data System (CO-RADS) [29], have been advocated to improve communication between radiologists and other health care providers by translating radiologic findings into standardized scores [29, 75, 76]. The CO-RADS scoring system assigns scores from 1 to 5 that increase with the level of suspicion of COVID-19 based on features seen on nonenhanced chest CT scans. Additionally, beyond assessing the likelihood of COVID-19, this and similar scoring systems also report on the extent of parenchymal involvement by assigning a CT severity score to patients highly suspected of having COVID-19 [77, 78]. Such standardized scoring systems enable fast and consistent clinical decision making, which is especially valuable in these enduring times of crisis [29, 75].

Artificial intelligence (AI) using deep learning has been advocated for automated reading of COVID-19 CT scans, including those used to diagnose COVID-19 [34, 35, 37, 38, 79–82] and quantify parenchymal involvement [42, 43, 83, 84]. Although these studies illustrate the potential of AI algorithms, their practical value is debatable [85]. Without adhering to radiologic reporting standards, it is doubtful that these algorithms provide any real benefit in addition to or instead of manual reading, limiting their adoption in daily practice. In addition, algorithms that follow a standardized scoring system need validation to confirm that they assign scores in a manner similar to that of radiologists and can be used to identify patients with COVID-19 with similar or even better performance.

The purpose of this study was to develop and validate an AI algorithm (the CO-RADS AI system) that automatically scores chest CT scans of patients suspected of having COVID-19 according to the CO-RADS and CT severity score systems. We compared findings of the CO-RADS AI system with readings of eight observers and with clinical assessments of the patients, including reverse transcription polymerase chain reaction (RT-PCR) test results.

2.2 Materials and methods

Medical ethics committee approval was obtained prior to the study. The need for written informed consent was waived, and data were collected and anonymized in accordance with local guidelines.

2.2.1 Study sample

We retrospectively included consecutive patients arriving at the emergency wards of an academic center and a large teaching hospital in the Netherlands in March and April of 2020 who underwent chest CT imaging for clinical suspicion of moderate to severe COVID-19. Criteria for CT were symptoms of lower respiratory tract infection, including cough, clinically relevant dyspnea requiring hospital admission, and fever with anosmia. CO-RADS and CT severity scores were reported as part of routine interpretation of the scans. Patients without scores in their radiologic report were excluded. Additionally, patients from the teaching hospital were excluded if they were known to have COVID-19 (proved with RT-PCR testing) prior to imaging or if RT-PCR test results were missing. The CT scanners were from different manufacturers; the protocols are described in Appendix 2.A.

Because RT-PCR testing may initially yield false-negative results, we considered patients to have COVID-19 if they had a positive RT-PCR test result or if their clinical presentation made COVID-19 the probable diagnosis. Criteria were the lack of an alternative diagnosis explaining the symptoms and admission to the intensive care unit due to respiratory failure, the need for high oxygen delivery, or unexplained death during admission.

Training and development set

The data of 476 patients from the academic center were used for model development, including 520 CT scans in total. CO-RADS scores were extracted from the radiologic reports. Scans with a CO-RADS score of 6, which signifies a positive RT-PCR result prior to imaging, were rescored independently by a chest radiologist with more than 30 years of experience (E.T.S.) who assigned CO-RADS scores of 1–5 to simulate an unknown RT-PCR status ($n = 52$). This observer was blinded to the original radiologic report and to all nonimaging data except age and sex.

Internal test set

A prior observer study assessing CO-RADS [29] reported on the remaining 105 patients included at the academic center. The data of these patients were set aside to verify the performance of the AI model. For each patient, at least one RT-PCR result was available within 5 days after CT scan acquisition. The earliest available scan of each patient was scored independently after CO-RADS classification by seven chest radiologists and one radiology resident (B.G., J.K., L.B., M.P., H.A.G., J.L.S., C. Schaefer-Prokop, T.v.R.V.) using a dedicated browser-based workstation (CIRRUS, Diagnostic Image Analysis Group, Nijmegen,

the Netherlands*). The observers were familiar with the CO-RADS and CT severity scoring systems from interpreting at least 30 scans. Four of them (B.G., J.K., J.L.S., and T.v.R.V.) had less than 5 years of experience in reading chest CT scans, and the others had up to 27 years. Any available later scans from these 105 patients were not used in this study. All observers were blinded to the RT-PCR test results and therefore could not assign a CO-RADS score of 6. Instead, they assigned CO-RADS scores of 1–5 based on their suspicion of COVID-19 pulmonary involvement. In addition, they semiquantitatively described the extent of parenchymal involvement per lobe using a predefined CT severity score on a six-point scale (0 = 0%, 1 = 1%–5%, 2 = 6%–25%, 3 = 26%–50%, 4 = 51%–75%, and 5 = >75%) [77].

External test set

The data of all patients included at the teaching hospital were set aside to verify the performance of the AI model on an external cohort. All these patients underwent RT-PCR testing on the same day as CT imaging. The CO-RADS score and the total CT severity score (the sum of the scores per lobe, as described previously) were extracted from the radiologic report of the earliest available scan of each patient.

2.2.2 Annotation of pulmonary lobes and opacities

Reference delineations of lung and lobar boundaries were automatically obtained for a convenience sample of 400 scans from the training and development set and for all 105 scans in the internal test set using commercial software (LungQ, version 1.1.1; Thirona, Nijmegen, the Netherlands), followed by manual correction. Reference delineations of areas with ground-glass opacities, consolidation, and mixed patterns were obtained for a convenience sample of 108 scans from the training and development set, as follows: Regions of parenchymal lung tissue with increased attenuation were identified with thresholding and morphologic operations. Vessels and airways were removed using automatic methods. Lesion candidates in lobes not affected by COVID-19 following the radiologic report were removed. The remaining lesion candidates were reviewed by a certified image analyst with at least 1 year of experience in correcting segmentations of pulmonary structures on chest CT scans. The analyst corrected the delineations and added and removed lesions, as needed.

*Available at `https://grand-challenge.org/reader-studies/`

2.2.3 Automated CT scoring

CT scans were scored automatically using three successively applied deep-learning algorithms. These performed pulmonary lobe segmentation and labeling, lesion segmentation and CT severity score prediction, and CO-RADS score prediction.

For lobe segmentation and labeling, we used a relational two-stage U-Net architecture specifically developed for robust pulmonary lobe segmentation [86]. The model was pretrained on 4000 chest CT scans from the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease study [87] and was fine-tuned with 400 scans from the present study.

For CT severity score prediction, we trained a three-dimensional U-Net using the nnU-Net framework [88] in a cross-validated fashion with 108 scans and corresponding reference delineations to segment ground-glass opacities and consolidation in the lungs. The CT severity score was derived from the segmentation results by computing the percentage of affected parenchymal tissue per lobe.

For CO-RADS score prediction, we used the three-dimensional inflated Inception architecture [89, 90], which is a three-dimensional extension of the state-of-the-art Inception image classification architecture [91]. The model was pretrained on ImageNet [92] and Kinetics [93] data sets and was trained with 368 CT scans from the present study to predict the corresponding CO-RADS score. The remaining scans from the training and development set were used to monitor the performance during training. Input to the model was the CT image together with areas of abnormal parenchymal lung tissue detected by the severity scoring algorithm.

Further details about the methods are provided in Appendix 2.B. The algorithm is freely accessible online[†].

2.2.4 Statistical analysis

Lobe segmentation was evaluated using the average Dice coefficient per lobe in the internal test set. Diagnostic performance of the automated CO-RADS scoring algorithm was evaluated using receiver operating characteristic curves and the area under the receiver operating characteristic curve (AUC). Youden index was used to determine the optimal threshold. Nonparametric bootstrapping with 1000 iterations was used to calculate 95% CIs. To quantify agreement, linearly weighted κ values and classification accuracy were determined by comparing the predicted CO-RADS and CT severity scores to the median of all combinations of scores from seven observers. The agreement of the AI system in terms of the linearly weighted κ value was compared with the agreement of the left-out observer

[†] https://grand-challenge.org/algorithms/corads-ai/

Table 2.1: Characteristics of training and test cohorts. Data are numbers with percentages in parentheses or means ± standard deviations. *p* values correspond to comparisons of the distribution of characteristics in the training set and the respective test set. *$p = 0.20$. †$p \leq 0.01$. ‡$p = 0.79$ for the internal test set, $p = 0.68$ for the external test set.

Parameter	Training set (*n* = 476)	Internal test set (*n* = 105)	External test set (*n* = 262)
Age	60 ± 16	62 ± 16*	64 ± 16†
Sex‡			
Male	267 (56)	61 (58)	154 (54)
Female	209 (44)	44 (42)	120 (46)
No. of CT scans per patient			
1	438 (92)	105 (100)	262 (100)
2	34 (7)		
3	2 (0)		
4	2 (0)		

using Monte Carlo permutation tests. CT severity scores were evaluated only for patients with a diagnosis of COVID-19. We tested for differences in demographic characteristics between training and test cohorts using *t* tests (age) and χ^2 tests (sex) and tested for differences in the sensitivity of the observers and algorithm at the specificity of the observers using McNemar tests. The significance level was .05. Analyses were performed with statistical software (R, version 3.6.2; R Foundation for Statistical Computing, Vienna, Austria) and Python, version 3.7.6 (scipy 1.5.0, sklearn 0.23.1, evalutils 0.2.3; Python Software Foundation, Wilmington, Del).

2.3 Results

2.3.1 Patient characteristics

A total of 581 and 262 consecutive patients were included at the academic center and the teaching hospital, respectively. The training and development set comprised 520 scans of 476 patients from the academic center, and the internal test set comprised 105 scans of the remaining 105 patients. The external test set comprised 262 scans of 262 patients from the teaching hospital. Six patients were excluded because CO-RADS scores were missing from their radiologic reports (Fig. 2.1). Patient characteristics for training and test sets are given in Table 2.1.

There were 58 patients with a clinical diagnosis of COVID-19 among the 105 patients in the internal test set (55%), and there were 179 such patients among the 262 patients in the

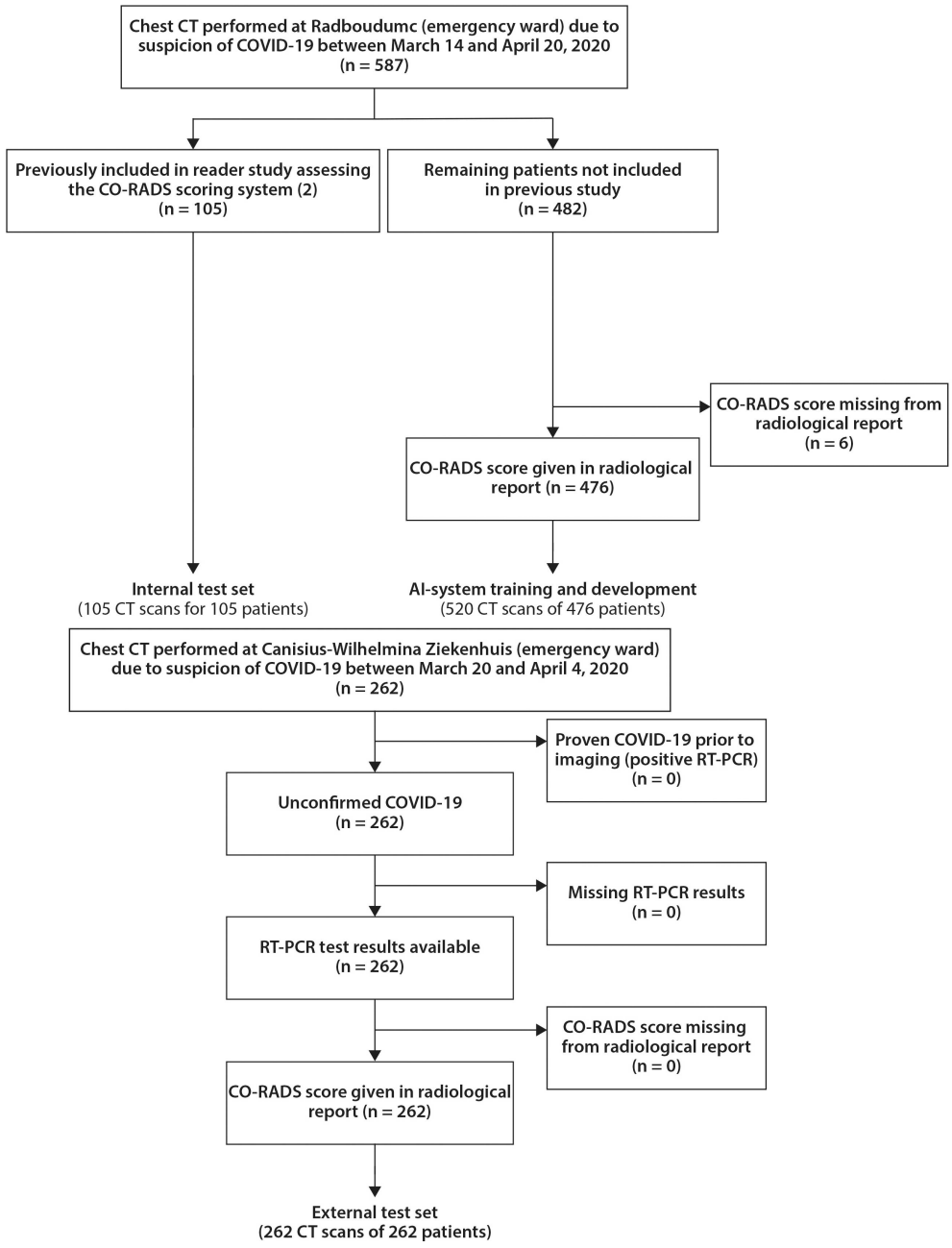


Figure 2.1: Flowchart shows patient inclusion in the training and test sets. Note that *n* refers to the number of patients. The number of CT images is higher in the training set, as several patients underwent multiple chest CT examinations during the inclusion period. However, in the test sets, only the earliest available scan for each patient is used. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019, RT-PCR = reverse transcription polymerase chain reaction.

Table 2.2: CO-RADS and CT severity scores according to radiologic reports or radiologist. The radiologist reviewed scans with an initial CO-RADS score of 6 (proven COVID-19). CO-RADS data are numbers of CT scans with percentages in parentheses. CT-severity-score data are medians and interquartile ranges and were reported only for cases with CO-RADS scores greater than or equal to 3. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019.

Parameter	Training set (*n* = 476)	Internal test set (*n* = 105)	External test set (*n* = 262)
CO-RADS categories			
CO-RADS 1	236 (45)	20 (19)	56 (21)
CO-RADS 2	66 (13)	10 (10)	12 (5)
CO-RADS 3	80 (15)	19 (18)	26 (10)
CO-RADS 4	45 (9)	17 (16)	25 (10)
CO-RADS 5	93 (18)	39 (37)	143 (54)
CT severity score	8 (9)	10 (7)	11 (7)

external test set (68%). Of these patients, 53 of 58 (91%) and 145 of 179 (81%) had a positive RT-PCR result, whereas the remaining patients diagnosed with COVID-19 had one or multiple negative RT-PCR test results but were diagnosed with COVID-19 on the basis of their symptoms. Table 2.2 summarizes the distribution of CO-RADS and CT severity scores according to the radiologic reports. The algorithm was executed successfully for all scans in the test sets. For the 105 scans in the internal test set, the median runtime of the algorithm was 212 seconds (range, 146–709), whereas the median reading time for radiologists was 82 seconds (range, 58–134 seconds).

2.3.2 Lobe segmentation

Reference delineations of lung and lobar boundaries were available for 104 of the 105 scans in the internal test set. In one image, the lobar boundaries could not be identified because of severe emphysema and the presence of an open window thoracostomy after a Clagett procedure in the right lung. In the remaining 104 images, the average Dice scores of the automatic lobe segmentations were 95.2% ± 2.0 for the left upper lobes, 92.4% ± 10.1 for the left lower lobes, 95.2% ± 3.1 for the right upper lobes, 92.2% ± 10.7 for the right middle lobes, and 94.7% ± 3.7 for the right lower lobes.

2.3.3 Identification of patients with COVID-19

In the internal test set, the algorithm distinguished between patients without and those with COVID-19 with an AUC of 0.95 (95% CI: 0.91, 0.98) on the basis of the probability of a CO-RADS score of 5 predicted by the algorithm. At the optimal threshold, the sensitivity of the algorithm was 85.7% (95% CI: 73.1, 98.2), and the specificity was 89.8% (95% CI: 79.6, 100).

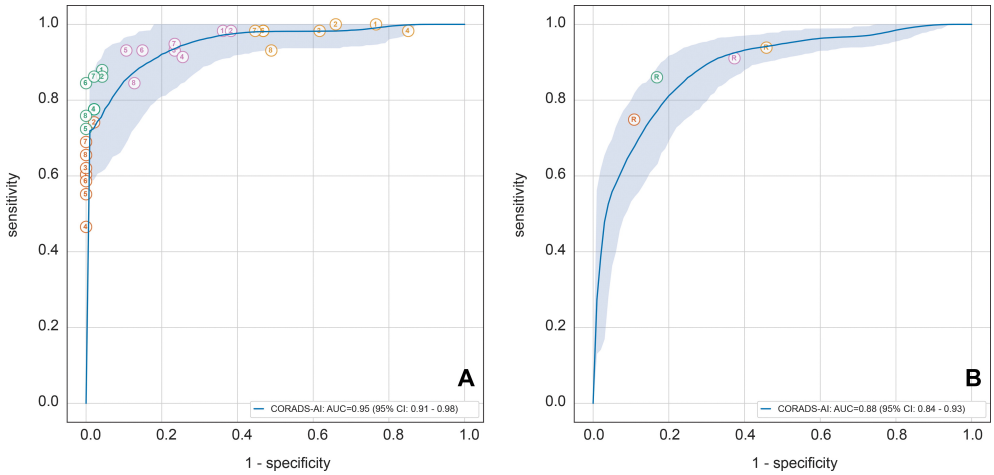


Figure 2.2: Receiver operating characteristic (ROC) curves for automatically predicted probability of a Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) score of 5 versus probability of a COVID-19 diagnosis. The ROC curve is based on the probability that the algorithm assigned a CO-RADS score of 5. The shaded area around the ROC curve reflects the 95% CI. **(A)** The performance of the eight observers is shown as individual points on the graph for the internal test set, and, **(B)**, the diagnostic performance of the scores from the radiologic reports is shown for the external test set. Different colors indicate different cutoffs, in which patients were considered to be predicted as having COVID-19 if the observer assigned a CO-RADS score of 5 (orange), 4 or 5 (green), 3–5 (magenta), or 2–5 (yellow). COVID-19 diagnosis meant either a positive reverse transcription polymerase chain reaction (RT-PCR) test result or very high clinical suspicion of COVID-19, despite at least one negative RT-PCR test result. AUC = area under the ROC curve, CORADS-AI = CO-RADS artificial intelligence system.

In the external test set, the AUC was 0.88 (95% CI: 0.84, 0.93) and sensitivity and specificity at the optimal threshold were 82% (95% CI: 69.7, 94.3) and 80.5% (95% CI: 67.9, 93.1), respectively. The corresponding receiver operating characteristic curves are shown in Figure 2.2, together with the operating points of the eight observers for the internal test set and with the operating points for the routinely reported CO-RADS scores for the external test set. In the internal test set, the mean sensitivity of the eight observers was 61.4% ± 7.9 (standard deviation) at a mean specificity of 99.7% ± 0.7, which was based on patients to whom they assigned a CO-RADS score of 5. In the external test set, the CO-RADS scores reported as part of clinical routine corresponded to a sensitivity of 134 of 179 (74.9%) and a specificity of 74 of 83 (89.2%). The sensitivities and specificities of the observers for each operating point and the sensitivities of the AI algorithm at the same specificities are given in Table 2.3. There was enough evidence for significantly better sensitivity of the observer for only three of the tested 36 operating points (8.3%) of all observers combined.

Table 2.3: Observer and AI sensitivity for identification of COVID-19 at specificity levels corresponding to various operating points of observers. Data are percentages or percentages, with 95% CIs in parentheses. *p* values correspond to comparisons of the sensitivity of the observer with that of the artificial intelligence (AI) algorithm. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019.

Parameter	Specificity	Observer sensitivity	AI sensitivity	*p* value
Internal test set (*n* = 105)				
CO-RADS 5				
Observer 1	100	60.3	71.9 (58.7, 84.8)	.143
Observer 2	97.9	74.1	73.0 (59.3, 88.3)	>.999
Observer 3	100	62.1	71.9 (58.7, 84.8)	.238
Observer 4	100	46.6	71.9 (58.7, 84.8)	<.001
Observer 5	100	55.2	71.9 (58.7, 84.8)	.064
Observer 6	100	58.6	71.9 (58.7, 84.8)	.143
Observer 7	100	69	71.9 (58.7, 84.8)	>.999
Observer 8	100	65.5	71.9 (58.7, 84.8)	.607
CO-RADS 4 + 5				
Observer 1	95.7	87.9	76.1 (61.2, 92.3)	.146
Observer 2	95.7	86.2	76.1 (61.2, 92.3)	.302
Observer 3	97.9	77.6	73.0 (59.3, 88.3)	.791
Observer 4	97.9	77.6	73.0 (59.3, 88.3)	.791
Observer 5	100	72.4	71.9 (58.7, 84.8)	>.999
Observer 6	100	84.5	71.9 (58.7, 84.8)	.118
Observer 7	97.9	86.2	73.0 (59.3, 88.3)	.180
Observer 8	100	75.9	71.9 (58.7, 84.8)	.815
CO-RADS 3 + 4 + 5				
Observer 1	63.8	98.3	97.2 (90.9, 100.0)	>.999
Observer 2	61.7	98.3	97.5 (91.4, 100.0)	>.999
Observer 3	76.6	93.1	93.6 (83.1, 100.0)	>.999
Observer 4	74.5	91.4	94.5 (84.7, 100.0)	.549
Observer 5	89.4	93.1	85.6 (69.2, 96.6)	.607
Observer 6	85.1	93.1	89.0 (75.8, 98.2)	.791
Observer 7	76.6	94.8	93.6 (83.1, 100.0)	>.999
Observer 8	87.2	87.3	87.3 (72.0, 96.9)	.815
CO-RADS ≥ 2				
Observer 1	23.4	100	99.1 (94.7, 100.0)	>.999
Observer 2	34	100	98.2 (93.8, 100.0)	>.999
Observer 3	38.3	98.3	98.2 (93.8, 100.0)	>.999
Observer 4	14.9	98.3	99.9 (98.1, 100.0)	.754
Observer 5	53.2	98.3	98.1 (93.1, 100.0)	>.999
Observer 6	53.2	98.3	98.1 (93.1, 100.0)	>.999
Observer 7	55.3	98.3	98.0 (92.9, 100.0)	>.999
Observer 8	51.1	93.1	98.1 (93.1, 100.0)	.503
External test set (*n* = 262)				
CO-RADS 5	89.2	74.9	67.6 (54.6, 81.7)	.035
CO-RADS 4 + 5	83.1	86	77.1 (63.1, 88.6)	.020
CO-RADS 3 + 4 + 5	62.7	91.1	92.5 (86.3, 96.7)	.585
CO-RADS ≥ 2	54.2	93.9	94.2 (90.1, 97.7)	>.999

Table 2.4: Agreement of observers and AI System with median score assigned by remaining seven observers in internal test set. Data are linearly weighted κ scores with 95% CIs in parentheses. *p* values correspond to comparisons of the κ scores of the left-out observer and AI system using the median of the other seven observers as the reference. AI = artificial intelligence, CO-RADS = Coronavirus Disease 2019 Reporting and Data System.

Parameter	Left-out observer vs. median observer	AI algorithm vs. median observer	*p* value
CO-RADS classification			
Observer 1	0.79 (0.73, 0.86)	0.59 (0.50, 0.70)	<.001
Observer 2	0.78 (0.71, 0.86)	0.59 (0.49, 0.70)	<.001
Observer 3	0.85 (0.80, 0.90)	0.60 (0.50, 0.69)	<.001
Observer 4	0.72 (0.65, 0.79)	0.60 (0.50, 0.71)	.018
Observer 5	0.74 (0.66, 0.81)	0.61 (0.51, 0.71)	.032
Observer 6	0.84 (0.79, 0.90)	0.61 (0.52, 0.71)	<.001
Observer 7	0.82 (0.77, 0.88)	0.58 (0.48, 0.68)	<.001
Observer 8	0.79 (0.73, 0.86)	0.61 (0.51, 0.70)	.002
Total CT severity score			
Observer 1	0.80 (0.74, 0.85)	0.53 (0.44, 0.63)	<.001
Observer 2	0.74 (0.68, 0.81)	0.53 (0.43, 0.62)	<.001
Observer 3	0.78 (0.71, 0.84)	0.54 (0.45, 0.63)	<.001
Observer 4	0.76 (0.69, 0.84)	0.53 (0.44, 0.63)	<.001
Observer 5	0.80 (0.74, 0.87)	0.54 (0.45, 0.63)	<.001
Observer 6	0.71 (0.64, 0.79)	0.55 (0.45, 0.64)	.004
Observer 7	0.81 (0.75, 0.87)	0.53 (0.44, 0.63)	<.001
Observer 8	0.73 (0.65, 0.80)	0.55 (0.46, 0.64)	.002

2.3.4 CO-RADS score prediction

When compared with the median CO-RADS score of all score combinations from seven of the eight readers of the internal test set, the automatically assigned CO-RADS score was in absolute agreement in 54.8% (460 of 8 × 105 = 840) of the patients and within one category of the seven observers' score in 80.5% (676 of 840). The remaining reader was in absolute agreement in 68.2% (573 of 840) of the patients and within one category of the other observers' score in 96.2% (808 of 840) of the patients. In the external test set, the AI algorithm score and the reference score were in absolute agreement in 64.12% (168 of 262) of the patients and within one category of one another in 85.50% (224 of 262) of the patients. The cross-tabulated results are given in Appendix Tables 2.C.1 and 2.C.2.

There was moderate to substantial agreement between AI-predicted CO-RADS scores and the observers' score according to the linearly weighted κ value (Table 2.4). For the internal test set, the mean (± standard deviation) κ value was 0.60 ± 0.01 for the AI system and 0.79 ± 0.04 for the left-out observer (p < .001 for all observers). For the external test set, the κ value was 0.69 (95% CI: 0.63, 0.75) for the AI system.

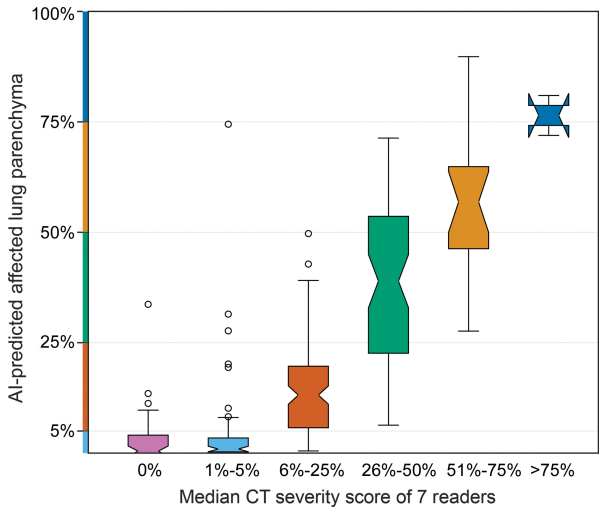


Figure 2.3: CT severity score predictions versus median of observer scores. The distribution of the percentage of affected lung parenchyma per lobe according to the automatic lesion (affected volume) and lobe segmentations (total volume) for the internal test set are shown as box plots. The notch in each box plot illustrates the 95% CI around the median. The CT severity score cutoffs are marked on the y-axis. AI = artificial intelligence.

2.3.5 CT severity score prediction

Because the automatic prediction is based on a segmentation of the lobes and abnormal regions in the lung, the algorithm outputs the percentage of affected parenchymal tissue rather than just the categorical severity score. Figure 2.3 depicts the percentage of affected parenchymal tissue per lobe with respect to the median severity score of the readers in the internal test set. The predicted score was in absolute agreement with the median score of all score combinations from seven of the eight observers in 17.2% (80 of 8 × 58 positive patients = 464) and deviated by not more than one point per lobe (*i.e.,* five points in total) in 94.0% (436 of 464 patients). In the external test set, the radiologic reports contained severity scores for 163 of the 179 patients with COVID-19. The AI algorithm was in absolute agreement with these scores in 17 of 163 (10.4%) patients and within one point per lobe in 146 of 163 (89.6%).

There was moderate agreement between AI-predicted severity scores and the observers' score according to the linearly weighted κ value (Table 2.4). For the internal test set, the mean κ value was 0.54 ± 0.01 for the AI system and 0.77 ± 0.03 for the left-out observer ($p < .001$ for all observers). For the external test set, the κ value was 0.49 (95% CI: 0.41, 0.56) for the AI system.

Representative examples of lobe segmentation results, CO-RADS score predictions, and CT severity score predictions with corresponding pulmonary lesions are shown in Figures

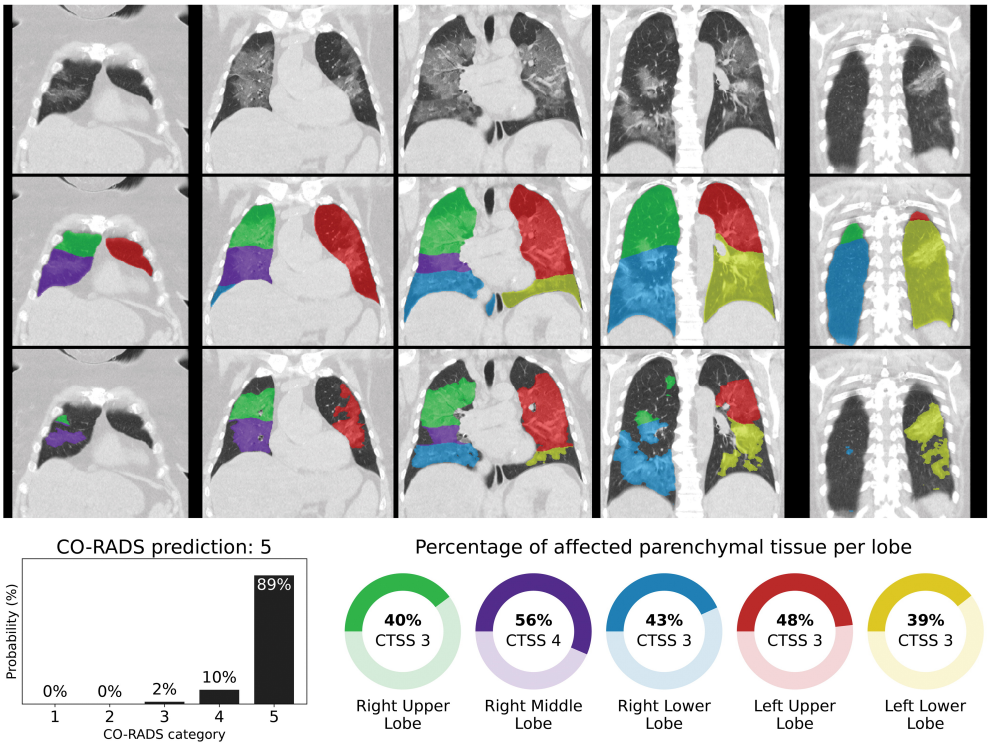


Figure 2.4: Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–positive case with extensive parenchymal involvement. Scans from a 73-year-old woman with a positive reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left) as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 3 (three observers), 4 (one observer), and 5 (four observers).

2.4, 2.5, 2.6, and Appendix 2.C.

2.4 Discussion

Artificial intelligence (AI) might be helpful in interpreting CT scans of patients highly suspected of having coronavirus disease 2019 (COVID-19), especially when it produces standardized output with which radiologists and other health care providers are familiar. In this study, we evaluated the performance of an AI system for automated scoring of chest CT scans of patients suspected of having COVID-19 on the basis of the COVID-19 Reporting and Data System (CO-RADS) and CT severity score classifications. This system identi-

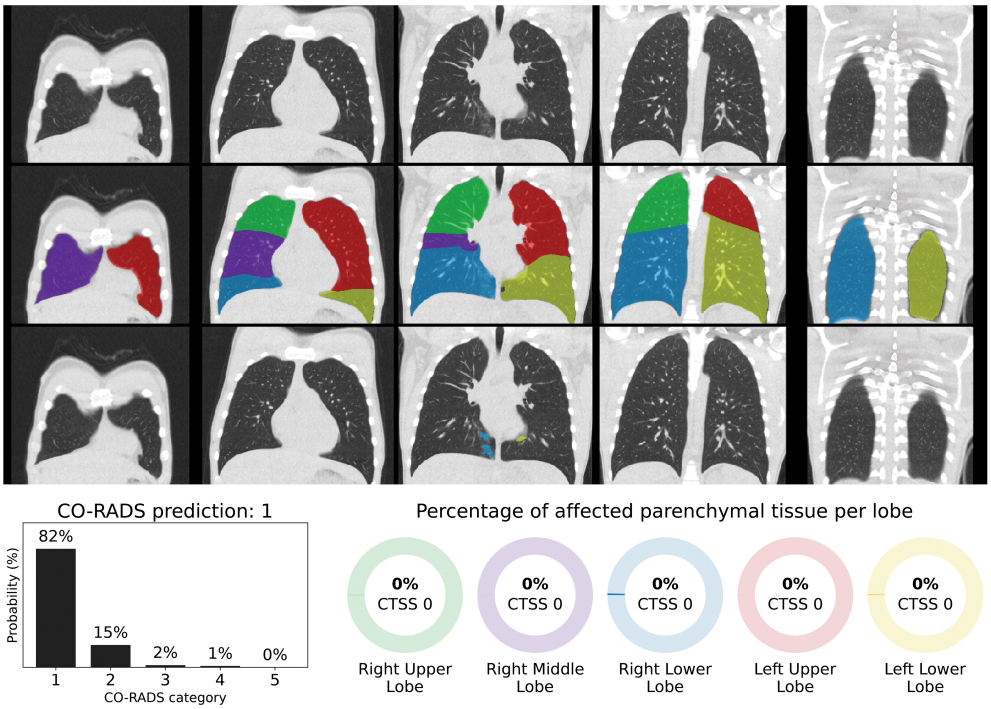


Figure 2.5: Coronavirus disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–positive case with little parenchymal involvement. Scans from an 18-year-old man with a positive reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 1 (two observers), 2 (five observers), and 3 (one observer).

fied patients with COVID-19 with high diagnostic performance, achieving an area under the curve (AUC) of 0.95 in the test set with sensitivity and specificity similar to those of the eight observers and achieving an AUC of 0.88 in the external test set. The automated CO-RADS scoring was in good agreement with that of the observers, although agreement was significantly higher between observers (mean κ value, 0.60 for AI agreement vs 0.79 for interobserver agreement; $p < .001$). Likewise, the automated CT severity score agreed well with observer scoring, with a mean κ value of 0.54 ± 0.01, which was also lower than the agreement among the observers ($\kappa = 0.77 \pm 0.03$, $p < .001$). An explanation may be that visually estimating the amount of affected lung parenchyma is subjective; studies have shown that human readers tend to overestimate the extent of disease [94]. In the four cases in which automatic measurements were more than 10 points higher than the reference, underlying causes were severe motion artifacts in three cases and opacifications caused by

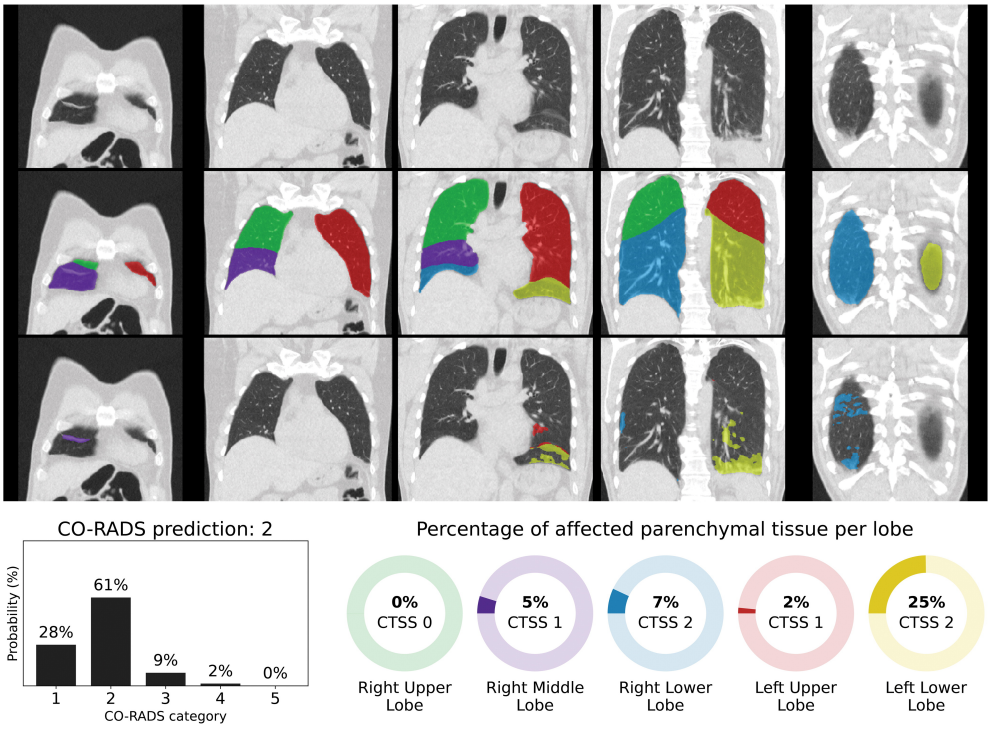


Figure 2.6: Coronavirus disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–negative case. Scans from a 54-year-old man with a negative reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 1 (three observers), 2 (three observers), and 3 (two observers).

aspiration pneumonia in one case. This emphasizes the importance of verification of automatically determined severity scores by humans.

In the short period since the outbreak, many groups have already developed AI algorithms for diagnosing COVID-19. Shi *et al.* [95] and Ito *et al.* [96] provide overviews of the proposed approaches. Most studies analyze small data sets and employ two-dimensional neural network architectures on axial sections. One of the first large studies by Li *et al.* [80] with such a two-dimensional approach showed performance on an independent test set comparable to ours but did not compare AI results with human reading. We experimented with their CovNet architecture but found that our three-dimensional, inflated Inception approach gave higher performance. Zhang *et al.* [84] also took a two-dimensional approach and showed in a large data set that segmentation of lesions per axial section

and feeding these slices into a classification network gave excellent results on several Chinese data sets, outperforming junior radiologists. Their method separates consolidations and ground-glass lesions, and this seems to be a promising approach. We followed the CT severity score that is part of CO-RADS and therefore have not separated different types of lesions. When tested on data from Ecuador, performance of their system decreased considerably. We also saw reduced performance on the external validation set for our method, but for that set, the CO-RADS scores obtained from routine practice also showed lower performance, comparable to the AI system. Similarly, Bai *et al.* [82] obtained good results with a two-dimensional approach (AUC = 0.95), reducing to an AUC of 0.90 on external data. To optimize AI software for data from different hospitals, Xu *et al.* [97] implemented a federated learning approach for COVID-19 diagnosis using CT scans and showed that this may lead to better performance on external data.

Mei *et al.* [42] developed a COVID-19 diagnostic AI system and tested it on a held-out set of 279 patients. They followed a two-dimensional approach similar to the methods discussed previously, but they added a separate neural network analyzing clinical symptoms and five laboratory findings. The combination of both networks provided the best performance (AUC = 0.92).

Given the overlap of morphologic features with other non-COVID-19–related diseases, our study could be advanced with an AI analysis of imaging, clinical signs, and routine laboratory parameters. There are other studies that have used visual scoring of imaging data and have demonstrated the potential of such a combined approach [98, 99].

None of the previously published works followed a standardized reporting scheme, and these systems therefore provide an unstandardized uncalibrated output. The lack of explicability of the output of AI systems is often seen as a limitation. By adhering to a standardized reporting system already validated in human observers [29], the proposed system overcomes this obstacle. The lobar output of the CT severity system is also familiar to radiologists. We did not find any other publications in which the accuracy of automated lobar segmentation and lesion segmentation on CT scans of patients suspected of having COVID-19 was quantitatively evaluated and compared with human readers.

Our study had several limitations. First, we trained the AI system with data from one medical center. More training is needed with multicenter data sets. Second, the study sample represented a population of patients who arrived in a high-prevalence situation at our hospital and underwent chest CT examinations because they were suspected of having COVID-19. There was only a limited number of patients with extensive preexisting lung disease in the training set. Third, inclusion of the study group took place after the influenza and respiratory syncytial virus season. Consequently, most CT images in this study were ei-

ther normal or demonstrated COVID-19 features. Ultimately, AI systems need to be trained with a larger data set before they can be expected to correctly interpret studies with overlapping abnormalities due to other types of pneumonia or other diseases, such as congestive heart failure, pulmonary fibrosis, or acute respiratory distress syndrome in patients without COVID-19.

In conclusion, our study demonstrates that an artificial intelligence (AI) system can identify patients with coronavirus disease 2019 (COVID-19) on the basis of unenhanced chest CT images with diagnostic performance comparable to that of radiologic observers. It is noteworthy that the algorithm was trained to adhere to the COVID-19 Reporting and Data System categories and is thus directly interpretable by radiologists. We believe the AI system may be of use to support radiologists in standardized CT reporting during busy periods. The automatically assessed CT severity scores per lobe may have prognostic information and could also be used to quantify lung damage during patient follow-up.

Chapter appendix

2.A CT protocol

At the academic center, scans were obtained on 160 or 320 multidetector CT (Canon Medical Systems, Otawara, Japan) without intravenous contrast injection, with a tube energy of 100, 120 or 135kV depending on body weight, collimation of 80 × 0.5 mm, rotation time of 0.275s, and were reconstructed using hybrid iterative reconstruction (AIDR 3D) with a preset noise tolerance of 35 into axial slices with 0.5 mm thickness and 0.5 mm increment. Median $CTDI_{vol}$ was 0.9mGy (IQR 0.6–1.6), and median DLP was 30.9 mGy*cm (IQR 18.2–50.3).

At the teaching hospital, scans were obtained on a 128 multidetector CT (Philips Medical Systems, Best, Netherlands) without intravenous contrast injection, with a tube energy of 100kV, collimation of 64 × 0.625 mm, rotation time of 0.4s, and were reconstructed using model-based iterative reconstruction (IMR1, SharpPlus kernel) into axial slices with 1 mm thickness and 1 mm increment. Median $CTDI_{vol}$ was 2.9mGy (IQR 1.8–4.2), and median DLP was 109.5 mGy*cm (IQR 72.3–162.0).

In both centers, a deep inspiration breath-hold technique was applied whenever feasible.

2.B Components of the AI system

The CORADS-AI system for automatic assessment of CO-RADS score and CT Severity Score in chest CT scans of patients with suspected COVID-19 is composed of three successively applied deep learning algorithms for (1) pulmonary lobe segmentation and labeling (2), CT severity score prediction and (3) CO-RADS score prediction. Following standard practice, we chose the architectures and hyper-parameters that obtained the highest accuracy on our development set (520 CT images). The test sets were not used for model development.

The first component of the CORADS-AI system automatically segments the pulmonary lobes in a CT scan using a deep learning algorithm based on a relational two-stage U-Net architecture (RTSU-Net) [86]. RTSU-Net is based on a cascade of two 3D U-Net architectures [100] and integrates a relational module to capture the global context of a large 3D CT volume using appearance and geometric information. This architecture takes a full CT

volume as input and outputs the corresponding segmentation mask of the five pulmonary lobes. The algorithm was first trained with 4000 chest CT scans from the COPDGene study [87] and then fine-tuned with 400 scans from the present study to increase the robustness to COVID-19 typical pulmonary abnormalities, such as consolidated areas. The remaining scans were used for validation during training, *i.e.,* for tuning the model parameters. Further details of the RTSU-Net architecture and training procedure can be found elsewhere [86].

Prior to executing the second and third components, which perform automated CO-RADS and CT severity scoring, the original CT volumes were resampled to an isotropic resolution of 1.5 mm × 1.5 mm × 1.5 mm voxel spacing to ensure that all images are interpreted at the same resolution. The volumes were cropped to the lungs based on the automatic lobe segmentation by cropping to the minimum bounding box of the union of all five lobes plus a 10 mm additional margin on each side. Since the CO-RADS model made use of transfer learning and therefore expected input values in the range 0 to 1, the images were additionally normalized for this model by clipping intensities below -1100 HU and above 300 HU and mapping the range 1100 HU to 300 HU to the range 0 to 1.

For CT severity score prediction, we trained a 3D U-net using the nnU-Net framework [88] with 108 scans in which we semiautomatically delineated ground-glass opacities and pulmonary consolidation[†]. The network was trained using a five-fold cross-validation scheme. The resulting probability maps were resampled to the original resolution of the CT scan and then thresholded at 50% probability. False-positive detections outside the lobes were automatically removed using the lobe segmentation mask. The volume of all detected lesions per lobe was divided by the volume of the corresponding lobe and the resulting percentage of affected parenchymal tissue was used to derive the CT severity score (0% = CTSS 0, 0%–5% = CTSS 1, 5%–25% = CTSS 2, 25%-50% = CTSS 3, 50%–75% = CTSS 4, > 75% = CTSS 5).

For CO-RADS score prediction, we used a deep learning model based on a 3D inflated Inception V1 architecture (I3D) [89, 90][§]. The I3D architecture builds upon the state-of-the-art Inception v1 model for 2D image classification but inflates the filters and pooling kernels into 3D. This enables the use of an image classification model pretrained with 2D data for a 3D image classification task. The model was pretrained in 2D with the ImageNet dataset [92], converted into a 3D model and then further pretrained using the Kinetics dataset [93], a 3D image classification dataset. The resulting pretrained 3D image classification model was trained with 368 scans of patients with a clinical suspicion of COVID-19 for CO-RADS score prediction, the remaining scans were used for validation during train-

[†] Source code available at `https://github.com/MIC-DKFZ/nnUNet`.
[§] Source code available at `https://github.com/deepmind/kinetics-i3d`.

ing (*i.e.,* for tuning the model parameters). Because the pretrained I3D models expect the 3D input data to have the same dimensions as the data the model was pretrained with, the cropped and isotropically resampled CT volumes were further cropped or padded to a fixed size of 240 × 240 pixels in-plane. In axial direction, 128 slices were uniformly sampled from the lung region. Lesion masks generated by the CT severity score model were preprocessed to the same shape and fed to the model together with the CT image. Since the I3D model expected a 3-channel input due to pretraining with RGB data, the first channel was the image, the second channel was the lesion mask, and the third channel was all zeros. All batch normalization layers in the model were fixed after pretraining. We trained the model with a batch size of 2, a learning rate of 0.0001 with Adam optimizer and cross entropy loss.

2.C Example output of the CORADS-AI system

We randomly selected two images for each CO-RADS score category (1 to 5) from the 105 CT scans in the internal test set, *i.e.,* images that the CORADS-AI system was not trained with, using the CO-RADS scores from the radiologic reports. The images, segmentation results and CORADS score and CTSS predictions are shown in Figures E1 to E5. Note that the CT severity score is intended for assessment of patients with clinical suspicion of COVID-19 who require hospitalization and is therefore typically not reported for unsuspicious cases. The radiologic reports of these 10 example cases therefore contained CT severity scores only for cases scored CO-RADS 3 or higher, these are reported in the figure captions.

Table 2.C.1: Scores of the AI-system cross-tabulated against the median scores of all combinations of seven observers for the internal test set. Data corresponds to the comparison of the AI-predicted CO-RADS scores with median scores of all leave-one-reader-out combinations, *i.e.*, 8 × 105 patients = 840 reference CO-RADS scores.

		Median of all combinations of seven of the eight observers				
		CO-RADS 1	CO-RADS 2	CO-RADS 3	CO-RADS 4	CO-RADS 5
AI	CO-RADS 1	64/840 (8%)	56/840 (7%)	0/840 (0%)	0/840 (0%)	0/840 (0%)
	CO-RADS 2	32/840 (4%)	64/840 (8%)	0/840 (0%)	0/840 (0%)	0/840 (0%)
	CO-RADS 3	24/840 (3%)	24/840 (3%)	24/840 (3%)	0/840 (0%)	16/840 (2%)
	CO-RADS 4	16/840 (2%)	16/840 (3%)	48/840 (6%)	24/840 (3%)	16/840 (2%)
	CO-RADS 5	0/840 (0%)	8/840 (1%)	72/840 (8%)	56/840 (7%)	280/840 (33%)

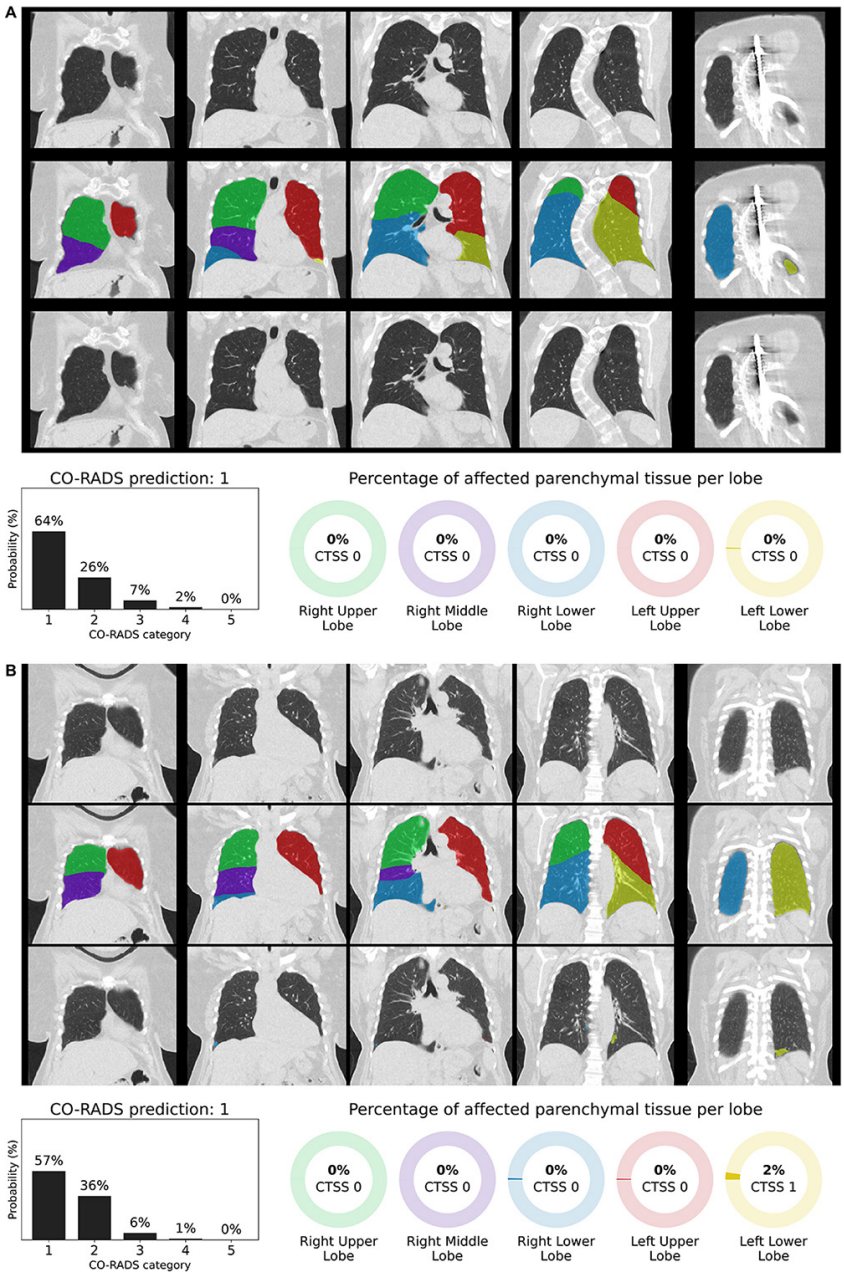


Figure 2.C.1: Model output for CO-RADS 1 cases according to the radiologic report. Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 8× CO-RADS 1 (top), and 3× CO-RADS 1, 4× CO-RADS 2 and 1× CO-RADS 3 (bottom).

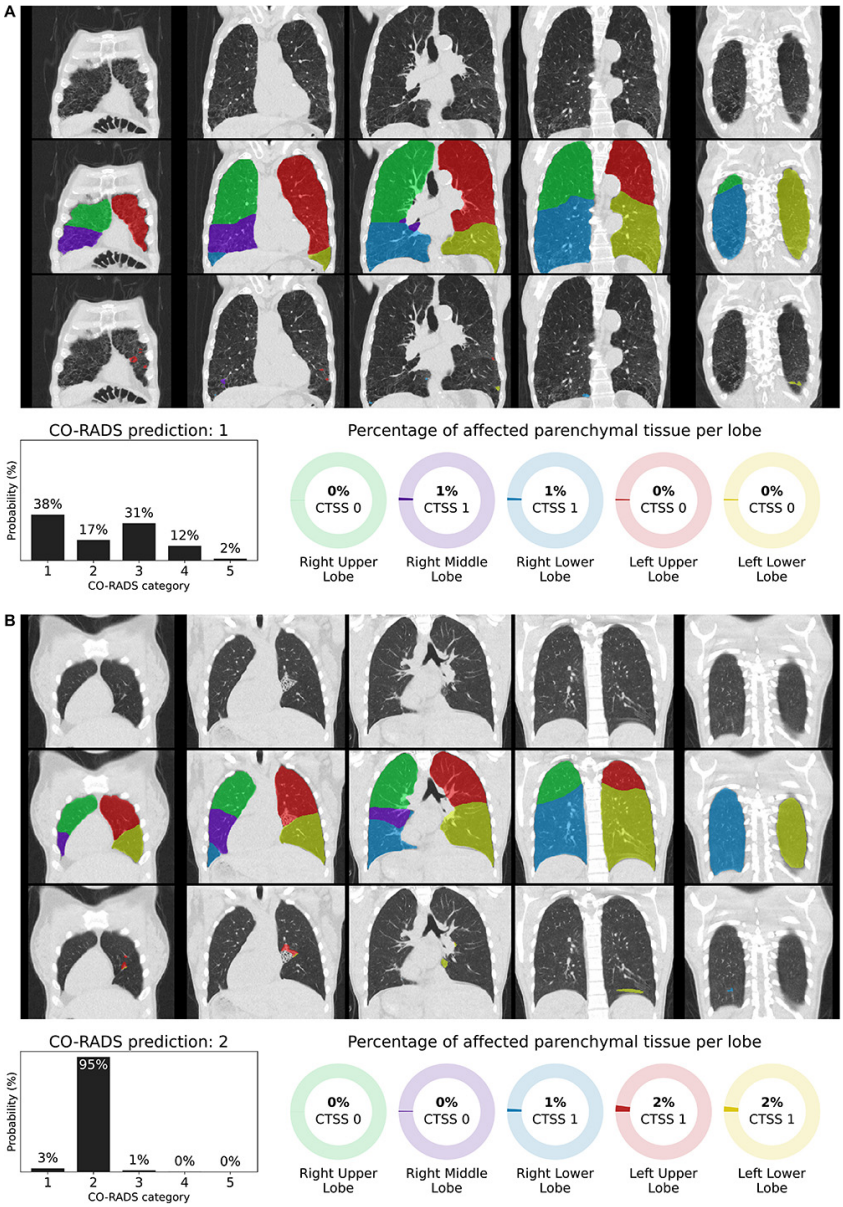


Figure 2.C.2: Model output for CO-RADS 2 cases according to the radiologic report. Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3× CO-RADS 1 and 5× CO-RADS 2 (top), and 4× CO-RADS 1, 1× CO-RADS 2 and 3× CO-RADS 3 (bottom). There was evidence of extensive emphysema with bilateral areas of ground glass (top) and of situs inversus (bottom).

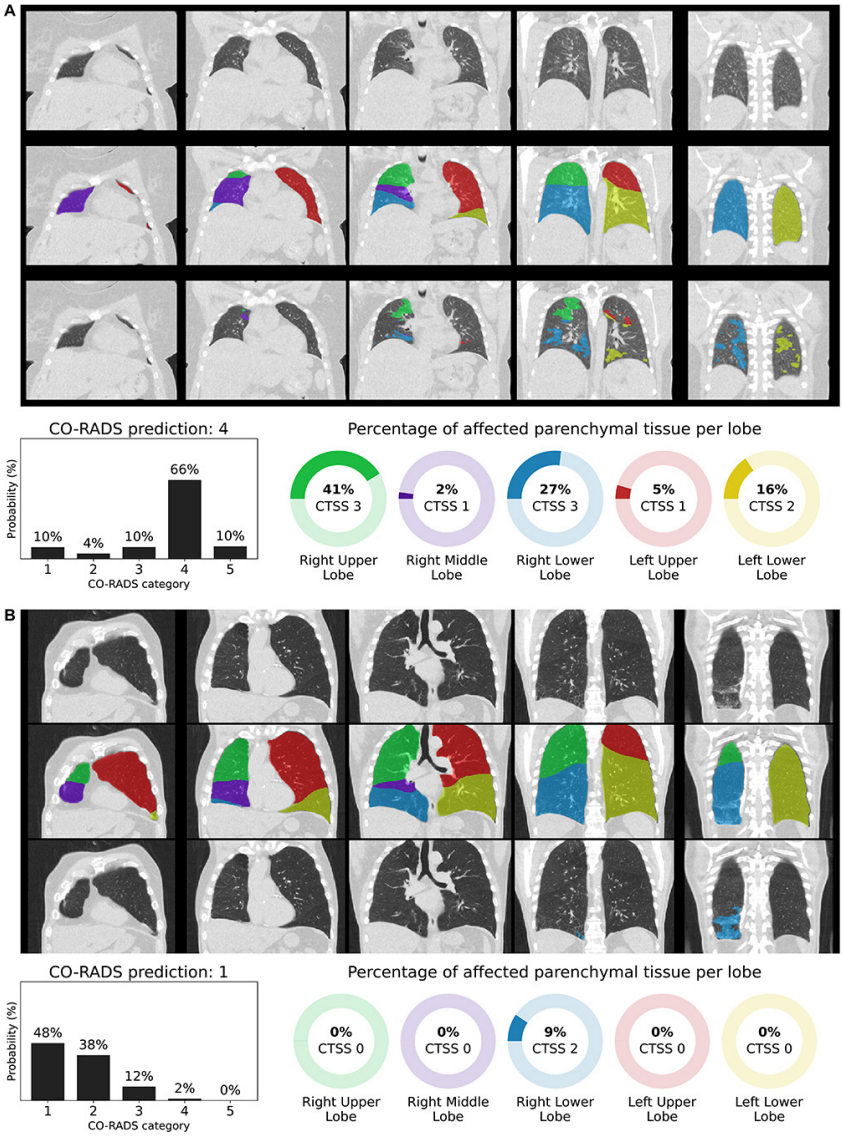


Figure 2.C.3: Model output for CO-RADS 3 cases with CTSS 12 (top) and CTSS 2 (bottom) according to the radiologic report Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3× CO-RADS 1, 2× CO-RADS 2 and 3× CO-RADS 3 (top), and 8× CO-RADS 2 (bottom). There was evidence of small ground glass areas in the peri-hilar regions (top) and patchy consolidation in the right lower lobe (bottom).

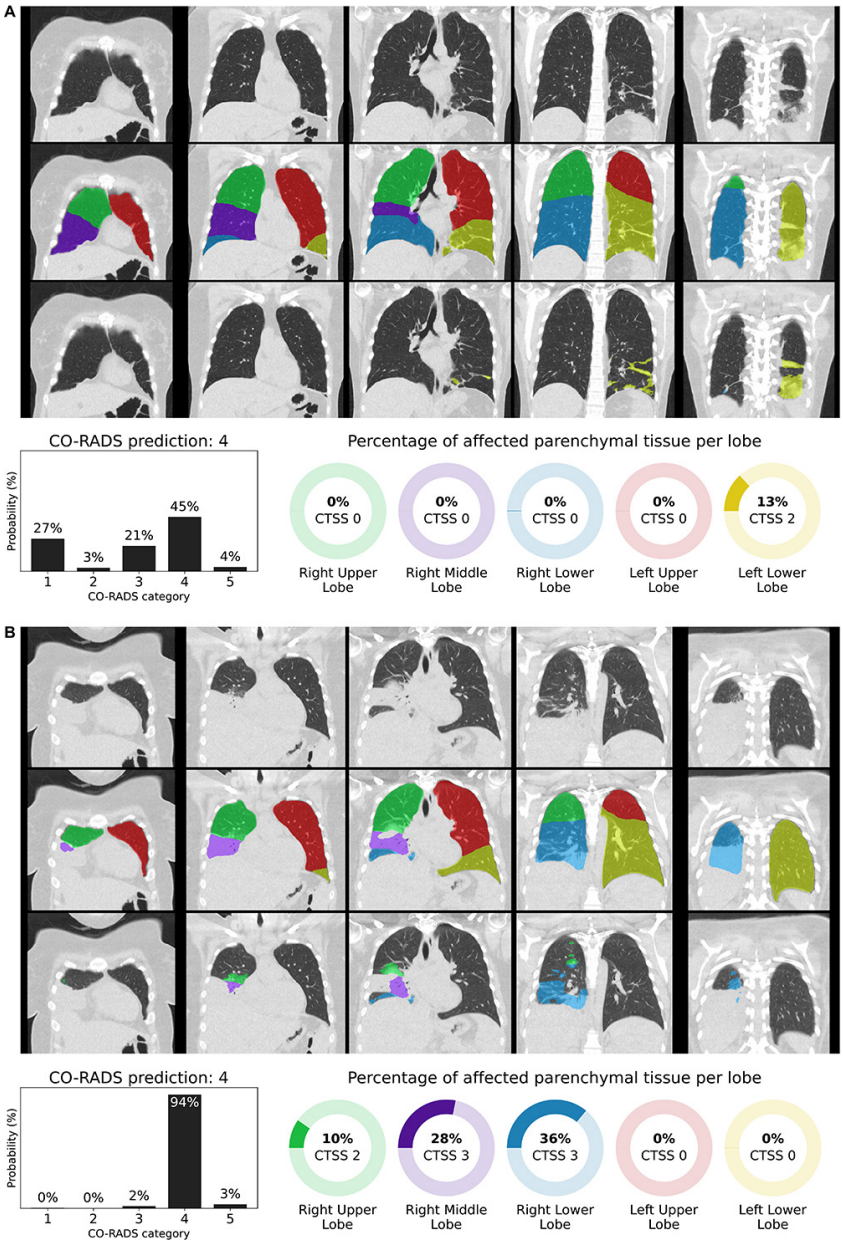


Figure 2.C.4: Model output for CO-RADS 4 cases with CTSS 5 (top) and CTSS 11 (bottom) according to the radiologic report Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3× CO-RADS 2, 4× CO-RADS 3 and 1× CO-RADS 4 (top), and 3× CO-RADS 3 and 5× CO-RADS 4 (bottom). There was evidence of ground glass areas in the periphery of the left lower lobe and areas of band-like collapse bilaterally (top), and of ground glass in the right upper lobe, dense consolidation in the entire middle lobe and posterior-basal in the right lower lobe (bottom).

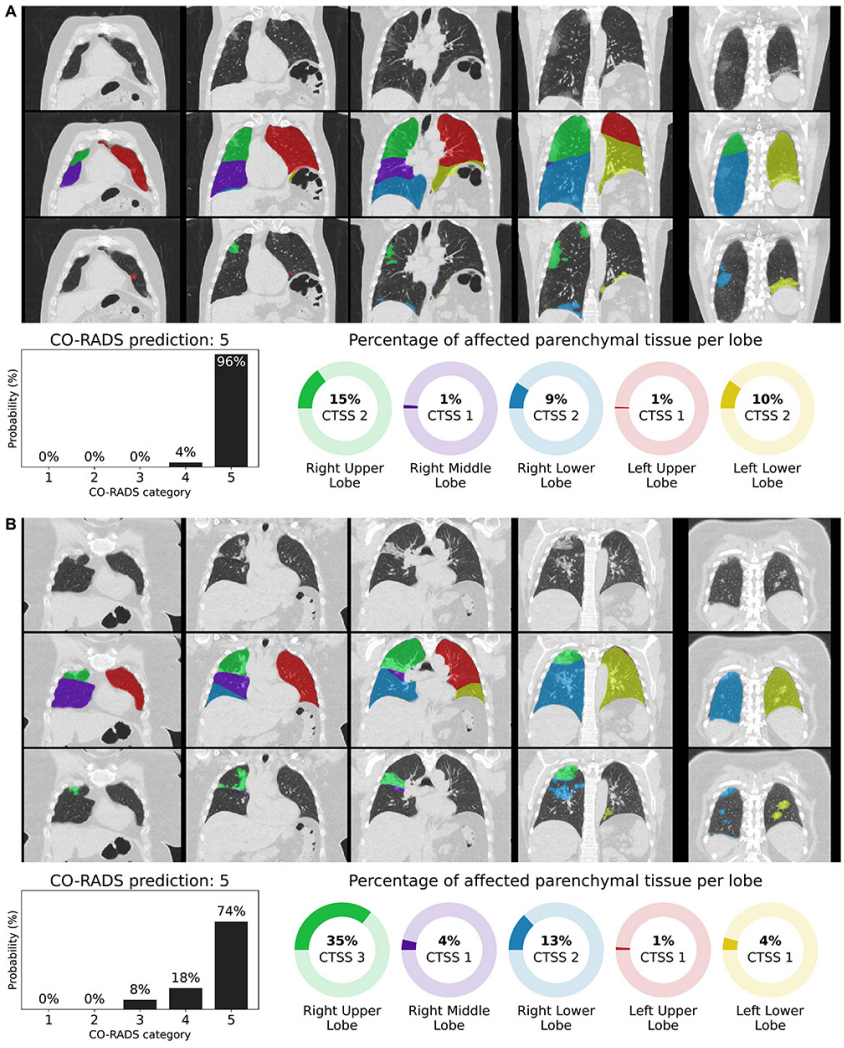


Figure 2.C.5: Model output for CO-RADS 5 cases with CTSS 8 (top) and CTSS 9 (bottom) according to the radiologic report Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3× CO-RADS 4 and 5× CO-RADS 5 (top), and 2× CO-RADS 4 and 6× CO-RADS 5 (bottom). There was evidence of areas of ground glass peripheral in the right lung, especially the upper lobe (top) and of extensive ground glass peripheral in the right lung as well as small areas in the left lung.

Table 2.C.2: Scores of the AI-system cross-tabulated against the scores from the radiologic reports of the scans in the external test set.

		Median of all combinations of seven of the eight observers				
		CO-RADS 1	CO-RADS 2	CO-RADS 3	CO-RADS 4	CO-RADS 5
AI	CO-RADS 1	37/262 (14%)	5/262 (2%)	7/262 (3%)	2/262 (1%)	0/262 (0%)
	CO-RADS 2	11/262 (4%)	2/262 (1%)	4/262 (2%)	1/262 (0%)	1/262 (0%)
	CO-RADS 3	7/262 (3%)	4/262 (2%)	8/262 (3%)	13/262 (5%)	14/262 (5%)
	CO-RADS 4	1/262 (0%)	0/262 (0%)	3/262 (1%)	5/262 (2%)	12/262 (5%)
	CO-RADS 5	0/262 (0%)	1/262 (0%)	4/262 (2%)	4/262 (2%)	116/262 (44%)

3

A systematic comparison of COVID-19 grading models

Authors: Coen de Vente*, Luuk H. Boulogne*, Kiran Vaidhya Venkadesh, Cheryl Sital, Nikolas Lessmann, Colin Jacobs, Clara I. Sánchez, Bram van Ginneken (*equal contribution)

Original title: Automated COVID-19 Grading with Convolutional Neural Networks in Computed Tomography Scans: A Systematic Comparison

Published in: IEEE Transactions on Artificial Intelligence (Volume: 3, Issue: 2, Pages: 129–138, April 2022)

https://doi.org/10.1109/TAI.2021.3115093

Abstract

Amidst the ongoing pandemic, the assessment of computed tomography (CT) images for COVID-19 presence can exceed the workload capacity of radiologists. Several studies addressed this issue by automating COVID-19 classification and grading from CT images with convolutional neural networks (CNNs). Many of these studies reported initial results of algorithms that were assembled from commonly used components. However, the choice of the components of these algorithms was often pragmatic rather than systematic and systems were not compared to each other across papers in a fair manner. We systematically investigated the effectiveness of using 3D CNNs instead of 2D CNNs for seven commonly used architectures, including DenseNet, Inception, and ResNet variants. For the architecture that performed best, we furthermore investigated the effect of initializing the network with pre-trained weights, providing automatically computed lesion maps as additional network input, and predicting a continuous instead of a categorical output. A 3D DenseNet-201 with these components achieved an area under the receiver operating characteristic curve (AUC) of 0.930 on our test set of 105 CT scans and an AUC of 0.919 on a publicly available set of 742 CT scans, a substantial improvement in comparison with a previously published 2D CNN. This paper provides insights into the performance benefits of various components for COVID-19 classification and grading systems. We have created a challenge on grand-challenge.org to allow for a fair comparison between the results of this and future research.

Impact statement: Applied artificial intelligence (AI) research focuses disproportionately on novel architecture modifications that do not necessarily generalize to other datasets, while neglecting systematic comparisons between commonly used algorithm components. This inhibits the deployment of AI for real-world applications. For automatic COVID-19 grading specifically, attention for compatibility of AI with clinical workflow is lacking. This paper presents a systematic investigation of COVID-19 grading algorithm components using a large publicly available dataset. The results are published in an online challenge. These contributions speed up the development of AI applications for COVID-19 grading by establishing insights into the components of such applications and by allowing applications produced by future research to be compared in a fair manner. The adherence to a standardized COVID-19 grading system may increase the compatibility between AI and clinical workflow. Altogether, this work may increase the efficiency and accuracy of radiologists when reading CT scans during this pandemic.

3.1 Introduction

Imaging of COVID-19 with chest computed tomography (CT) has been found to be helpful for diagnosis of this disease in the current pandemic [30]. With the aim to reduce the workload of radiologists, various machine learning techniques have been proposed to automatically grade and classify the presence of COVID-19 in CT images [31–46, 101–106]. Automatic COVID-19 classification methods have already been deployed in several medical centers [36].

By far the most common technique for automatic COVID-19 classification from CT images is the Convolutional Neural Network (CNN) [107, 108], which is the current state-of-the-art for image classification [4]. The works that use this approach can be divided into those that use 2D CNNs [31, 32, 34, 35, 41–45, 103] and those that use 3D CNNs [33, 37–41, 46, 104, 105]. While 3D CNNs are directly capable of exploiting 3D information present in CT volumes, 2D CNNs can only indirectly use 3D information by aggregating their output for individual slices of the image to produce an image level prediction. 3D CNNs are typ-

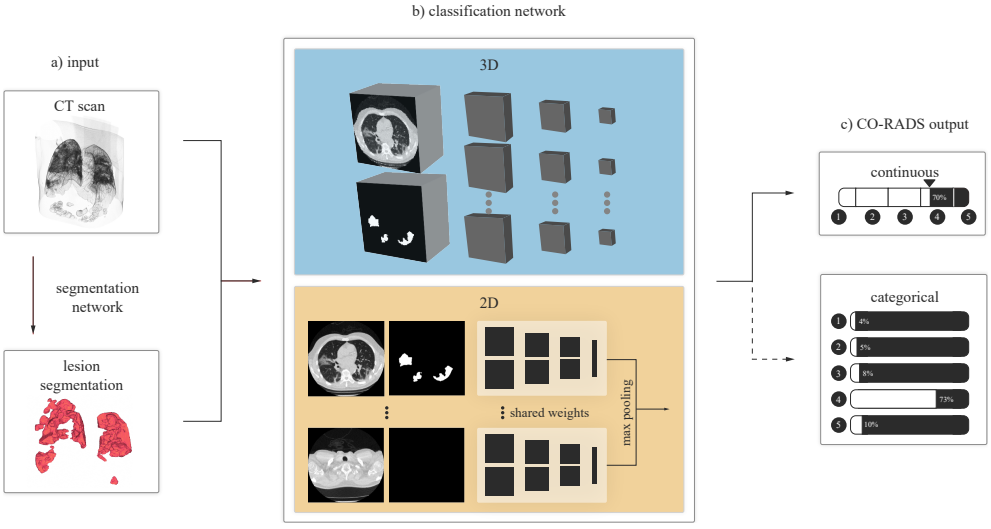


Figure 3.1: Schematic representation of the different components used for CO-RADS grading from CT scans using convolutional neural networks in patients suspected with COVID-19. This processing pipeline was used in all experiments of this work. **(a)** The input CT scan is fed into a lesion segmentation network. The CT and the lesion segmentation are used as separate input channels to the classification network as described in Section 3.3.3. In one of the ablation study experiments, this lesion segmentation input was left out. **(b)** We compared a variety of 3D (top) and 2D architectures (bottom) as described in Section 3.3.3. The 3D architectures take as input the full volume. The 2D architectures use individual slices as input. **(c)** We compared a continuous output to a categorical output in the ablation study. Section 3.3.3 describes the continuous output in detail. The dashed line indicates that the categorical output replaces the continuous output in one of the models in the ablation study and all models in the architecture search, but it is not incorporated in the main approach.

ically more memory intensive than 2D CNNs, but Graphics Processing Units (GPUs) with sufficient memory to train 3D models are becoming increasingly available. Moreover, radiologists are specifically instructed to take 3D information into account by inspecting different orthogonal views for assessing the suspicion of COVID-19 in CT scans [29]. This indicates that 3D information is essential for radiologists in assessing the patterns indicative for COVID-19. Additionally, the slice thickness of CT scans are increasingly becoming smaller [109] so that the scans contain more detailed 3D information. We therefore hypothesize that 3D CNNs are more suitable for COVID-19 classification from CT scans than 2D CNNs.

A major issue that inhibits the utilization of artificial intelligence in real-world applications, such as COVID-19 diagnosis from CT, is the excessive focus of research on novel architectures, while scientifically sound comparisons and proper evaluations on external datasets are lacking. Often, small additions and adaptations to model architectures for incremental improvements on specific datasets are proposed that do not generalize well to other datasets. This issue is increasingly being recognized and simple baselines have been proposed which perform comparably to or better than over-engineered solutions [26, 27].

The goal of this paper is therefore not to introduce novel architectural tweaks, but instead to perform a comparative study that evaluates existing approaches. To indicate the generalization capabilities of automatic COVID-19 classification systems, some methods have been validated on data from different centers than the data that were used for training [33, 40]. Also, the same validation methods, such as receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC), have been reported across different studies [31–46]. However, since each study used different datasets for training and for validation, the need for fair, direct comparisons of the performance of these algorithms remains unsatisfied. Recently, the "CT images and clinical features for COVID-19" (iCTCF) dataset was made publicly available [110], enabling a fair comparison of COVID-19 classification methods.

This paper compares a variety of 2D and 3D CNN architectures for COVID-19 classification. We trained and evaluated the approaches on the same internal dataset. Moreover, in an ablation study, we investigated performance changes due to 1) using transfer learning for 2D and 3D COVID-19 classification models, 2) using prior information in the form of COVID-19 related lesion segmentations as additional input to the network, 3) replacing the categorical output with a continuous output.

We furthermore created a public challenge [111] for evaluating and comparing different COVID-19 classification algorithms. Algorithms can be submitted to the challenge as Docker containers and are evaluated on the iCTCF dataset that we used in this paper. This

allows their performance to be compared to the methods presented in this paper, as well as to other COVID-19 grading and classification algorithms that are submitted to the challenge.

3.2 Background

3D CNNs were initially proposed for processing video data [108], where the third dimension of the convolutional layers dealt with the temporal dimension. In later works, 3D CNN architectures were derived from 2D CNN architectures by expanding the 2D filters into 3D [89]. Methods based on these inflated 3D CNNs, in particular the Inflated Inception-v1 (I3D) model, have recently been successfully employed for lung nodule detection and scan-level classification tasks from thorax CT scans [90, 112].

The large majority of the architectures used for COVID-19 classification from CT scans in previous works [31–40, 42–46, 102, 104, 105, 113] are heavily or completely based on the ResNet [114], DenseNet [115], or Inception [91] architecture families. Especially ResNet architectures have been used frequently [31, 32, 34, 36–38, 42–44, 104, 105, 113]. Some works did not use a full ResNet architecture, but did incorporate residual blocks into their model [45, 102]. Architectures from the DenseNet [31, 33, 46] and Inception [35, 40] families have been used less frequently. Other architectures such as VGG-19 [6], Inception-ResNet-v2 [116], NASNet [117], and EfficientNet [118] have also been used in research for COVID-19 classification from CT scans [82, 113, 119–121]. Due to the lack of standardized data for testing across different works, previous research does not identify which architecture produces the best performance for COVID-19 classification from CT.

Fine-tuning is a widely used technique in research on deep learning in medical imaging [122] and COVID-19 classification specifically [123]. With fine-tuning, models are initialized with pre-trained weights from models trained on a different task or dataset. They are commonly pre-trained on the ImageNet [124] dataset that contains a large variety of 2D natural images. Afterwards, the models are trained for the task at hand. Pre-training speeds up training and can offer performance gains for large models [122]. It has been used in several 2D CNN COVID-19 classification methods [32, 34, 35, 43, 44]. Pre-trained weights have also been used for 3D CNN-based methods. Wang *et al.* [33] pre-trained their model for COVID-19 classification on a large number of CT scans from lung cancer patients. Inflated 3D CNNs can conveniently be initialized by inflating 2D weights. 2D weights have been used to pre-train I3D models for video classification [89] and chest CT classification [90] tasks.

Before presenting CT images to the CNN, they are often pre-processed by extracting

the lung region using lung or lobe segmentation algorithms. These lung regions are then used either for cropping around and centering to the lungs [33, 34, 40, 43, 104] and/or by suppressing non-lung tissue [32–34, 36–39, 42, 105]. Yang *et al.* [31] used a lung segmentation as an additional input channel and used lesion masks as extra information by training their model to perform lesion segmentation and COVID-19 classification simultaneously. Lessmann *et al.* [40] also added a lesion segmentation to the input of their model.

Most studies on automated detection of COVID-19 employ a categorical classification output format that uses a softmax or sigmoid activation [123]. Previous works have trained models to discern between COVID-19 positive and negative patients [31, 34, 39, 42–46, 101, 102, 104], COVID-19 positive patients and patients with other types of pneumonia [33, 35, 37], and between all three [32, 38, 105]. In this work, we followed Lessmann *et al.* [40] and trained our models to produce CO-RADS [29] scores on chest CT scans of suspected COVID-19 patients. The CO-RADS score denotes the suspicion of COVID-19 on a scale from 1 to 5 and was developed to standardize reporting of CT scans of patients suspected with COVID-19 [29]. Scoring systems, like CO-RADS, have been advocated for better communication between radiologists and other healthcare providers [29, 40].

3.3 Methodology

3.3.1 Data

Training and internal test data

The internal dataset contained CT scans from consecutive patients who presented at the emergency wards of the Radboud University Medical Center, the Netherlands in March, April and May 2020 and were referred for CT imaging because of moderate to severe COVID-19 suspicion. The retrospective and anonymous collection of this data was approved by the ethical review board of Radboudumc (CMO2016-3045, Project 20027) prior to the study. Further details such as imaging parameters can be found elsewhere [40].

CO-RADS scores were reported by a radiologist as part of routine interpretation of the scans. CO-RADS 1 was used for normal or non-infectious etiologies, having a very low level of suspicion. CO-RADS 2 was used if the CT-scan was typical for other infections than COVID-19, indicating a low level of COVID-19 suspicion. CO-RADS 3 implies equivocal findings and features compatible with COVID-19, but characteristics of other diseases are also found. CO-RADS 4 and 5 indicate a high and very high level of COVID-19 suspicion, respectively.

We randomly split the dataset into a development set with 616 patients and an internal

Table 3.1: Number of CT images in internal dataset.

	CO-RADS							
	1	2	3	4	5	Total	Neg	Pos
Development set								
Training	253	71	78	37	73	512	324	188
Validation	81	24	26	11	23	165	105	60
Internal test set	20	10	19	17	39	105	30	75
Total	354	105	123	65	135	782	459	323

Table 3.2: Number of CT images in external dataset.

Grade [41]							
Control	Mild	Regular	Severe	Critically ill	Total	Neg	Pos
207	23	363	117	32	742	207	535

test set of 105 patients. The patients in the development set were split into 75% for training and 25% for validation using data stratification based on the CO-RADS scores. The distribution of CO-RADS scores over the different splits is displayed in Table 3.1. All data splits were made such that all scans from a patient with multiple visits ended up in the same split.

External test data

For external evaluation, we used the publicly available CT images and clinical features for COVID-19 dataset (iCTCF) dataset [41, 110]. Since we focused on comparing architectures for CT image processing for COVID-19 classification, we did not incorporate the clinical features from this dataset into the input for our models. In iCTCF, patients were categorized with a Chinese grading system that distinguishes the classes as Control, Mild, Regular, Severe, Critically ill and Suspected. Since there was no etiological evidence available for the presence of COVID-19 in Suspected cases [41], we did not use them for testing our models. The distribution of the other classes is displayed in Table 3.2. The grading system uses etiological laboratory confirmation and other factors such as clinical features and CT imaging [41]. The control cases include both healthy patients and patients with community acquired pneumonia. Most of the iCTCF data has been made publicly available, but some CT scans were not available at the time of conducting this study. We validated our models with all available data from the first iCTCF cohort for which etiological evidence for the presence of COVID-19 was available [110].

3.3.2 2D and 3D architectures

We compared the performance of a variety of popular 2D and 3D CNN architectures for the task of COVID-19 classification from CT. More specifically, we compared vanilla 2D and 3D versions of DenseNet-121, DenseNet-169, DenseNet-201, Inception-v1, ResNet-18, ResNet-34, and ResNet-50. Section 3.2 describes previous works that have used many of these architectures.

Since we used scan-level labels for training and testing these models, the 2D architectures required the integration of a slice-wise reduction step, while the 3D architecture did not. For the 2D architectures, we therefore integrated the slice-wise reduction step presented by Li *et al.* [32]. First, the 2D CNN extracts features of individual axial slices. A global max pooling step reduces these features to a 1D vector, to which a fully connected layer is applied with an output size equal to the number of classes.

3.3.3 Ablation study

We investigated whether additional model components had an effect on COVID-19 classification performance in an ablation study. Fig. 3.1 shows a summary of the processing pipeline that was used.

Since performing the ablation study for all 2D and 3D architectures would require a large quantity of computational resources, the ablation study was instead performed with only the best performing architecture in terms of quadratic weighted kappa (QWK).

Lesion map as prior information

To aid the model in localizing COVID-19 related parenchymal lesions, we provided a lesion segmentation map as additional input in a separate input channel. More specifically, the CT image was fed into the first input channel, the lesion segmentation into the second channel, and the third channel was presented with zeros. When training models without the additional lesion segmentation input, the CT image was fed into all three input channels.

A 3D nnU-Net [27] trained by Lessmann *et al.* [40], which segments ground-glass opacities (GGOs) and consolidations, provided the lesion segmentations. GGOs and consolidations are biomarkers with major importance in diagnosing COVID-19 [29].

Dimensionality

Since various components were added to the models in the ablation study, we trained both the 2D and 3D variants of the best performing architecture. This allows for an analysis of the performance difference solely due to the dimensionality of the model in our complete processing pipeline.

Pre-training

We investigated the performance changes due to pre-training on a natural image classification task. The 2D models were initialized with weights pre-trained on ImageNet. The 3D models were initialized with the same weights by inflating the pre-trained 2D convolution kernels to 3D.

Continuous output

The standard output format of CNNs used for categorical classification does not capture the ordinal nature of the CO-RADS scoring system. Furthermore, although the CO-RADS scoring system allows for a higher level of interpretability than a binary system, the fact that a CO-RADS suspicion score of 3 indicates that it is unclear whether COVID-19 is present makes it difficult to decide on the onset of the positive class for the predicted scores in ROC analyses. For these reasons, we considered the CO-RADS classification to be a regression task. Hence, the model had one output node that was forced to the range (0, 1) using the sigmoid function. CO-RADS scores were mapped to target values in the range [0, 1] with a uniform spacing between CO-RADS classes such that CO-RADS scores of 1 and 5 were assigned target values of 0 and 1, respectively. As the network had one output node, binary cross-entropy was used as loss function. With this method, unlike a standard categorical approach with a softmax layer and categorical cross-entropy loss, predictions that are further off from the target are penalized more heavily than predictions that are closer. To obtain a CO-RADS score during inference, the sigmoid output was multiplied by 4, rounded to the nearest integer and added to 1. De Vente *et al.* [125] explored this approach for prostate cancer grading and found that it outperformed other regression and categorical output methods.

3.3.4 Pre-processing

The CT scans were clipped between -1100 and 300 Hounsfield units, normalized between 0 and 1, and resampled to a voxel spacing of 1.5 mm^3 using linear interpolation. The scans

were further pre-processed using a lung segmentation algorithm that was trained on data from patients with and without COVID-19 [86]. More specifically, any slices with a distance of 10 mm or more to the lung mask were discarded and the remaining slices were cropped to 240 × 240 pixels around the center of the mask. Following previous research with I3D models [89, 90, 112], we trained our models with a fixed 3D input size. To achieve this without adding extra slices that do not contain information regarding the presence of COVID-19, we uniformly sampled 128 axial slices along the z-axis.

3.3.5 Training

We trained all networks with a batch size of 2, the Adam optimizer with β_1 = 0.9, β_2 = 0.999, and a learning rate of 10^{-4}. Data augmentation consisted of random zooming between -20% and +20%, rotation between -15% and +15%, shearing between -10% and +10% and elastic deformations in the axial plane, translation between -2 and +2 voxels in the z-direction, -20 and +20 voxels in both the x- and y-direction, and additive Gaussian noise with a mean of 0 and a standard deviation between 0 and 0.01 (after intensity normalization between 0 and 1). To correct for the class imbalance, we monitored the performance on the validation data in the development set during training with balanced samples based on the distribution of CO-RADS classes in the training set. We used early stopping with a patience of 10 000 training batches and the QWK on the validation set for the stopping criterion. Gradient checkpointing [126] reduces GPU memory requirements for training deep neural networks without affecting performance. This technique was used when necessary to enable a batch size of 2 for the 2D models.

To rule out the possibility that performance differences between the 3D and 2D approach were due to other factors such as pre-processing or data augmentation, we kept all hyperparameters the same during training.

Each model was trained on a single GPU, using NVIDIA GeForce GTX TITAN X, GeForce GTX 1080, GeForce GTX 1080 Ti, GeForce RTX 2080 Ti, TITAN Xp, and A100 SXM4 cards.

3.3.6 Ensembling

The models were sensitive to the randomness of the training process introduced by initialization of weights without pre-training, sample selection, and data augmentation. In order to enable stable comparisons, we obtained ensembles by training 10 instances of the same model with different random seeds. The ensemble output was obtained by simply taking the mean of the individual model outputs. For categorical model ensembles, the output was the mean of the probability output vectors of the individual models. All results

Table 3.3: Architecture properties.

Dim.	Architecture	Parameter count ($\times 10^6$)	Inference time (ms)	FLOP count ($\times 10^{11}$)
2D	DenseNet-121	6.88	151.09 ± 8.76	8.43
	DenseNet-169	12.33	255.21 ± 17.47	9.93
	DenseNet-201	17.87	326.30 ± 3.81	12.70
	Inception-v1	5.59	40.92 ± 10.16	4.47
	ResNet-18	11.17	8.95 ± 1.25	5.52
	ResNet-34	21.27	13.53 ± 1.42	11.06
	ResNet-50	23.47	35.91 ± 10.50	12.39
3D	DenseNet-121	11.24	25.07 ± 8.49	10.88
	DenseNet-169	18.54	31.49 ± 11.48	11.30
	DenseNet-201	25.33	38.48 ± 15.65	12.14
	Inception-v1	12.29	36.74 ± 16.75	5.13
	ResNet-18	33.21	28.33 ± 31.31	6.08
	ResNet-34	63.52	22.56 ± 14.37	9.29
	ResNet-50	46.21	31.09 ± 8.49	7.39

presented in Section 3.4 were obtained from ensembles unless stated otherwise.

3.3.7 Evaluation

We evaluated the CO-RADS scoring performance using the QWK score. This measure accounts for the ordinal nature of the CO-RADS score by weighting mismatches between true and predicted labels differently based on the magnitude of the error. Following previous works on COVID-19 classification and grading [32–41], diagnostic performance was evaluated using the AUC and ROC curves.

We calculated 95% confidence intervals (CIs) with non-parametric bootstrapping and 1000 iterations [127]. Statistical significance was computed with the same bootstrapping method [128].

The AUCs that our models achieved on the external test set are additionally listed on the Grand Challenge platform [111] to allow for a direct comparison between our and future COVID-19 grading and classification solutions.

Inference duration was calculated on the same machine for each architecture, using a GeForce RTX 2080 Ti card. The reported durations were averaged over 50 forward passes of a batch with one sample.

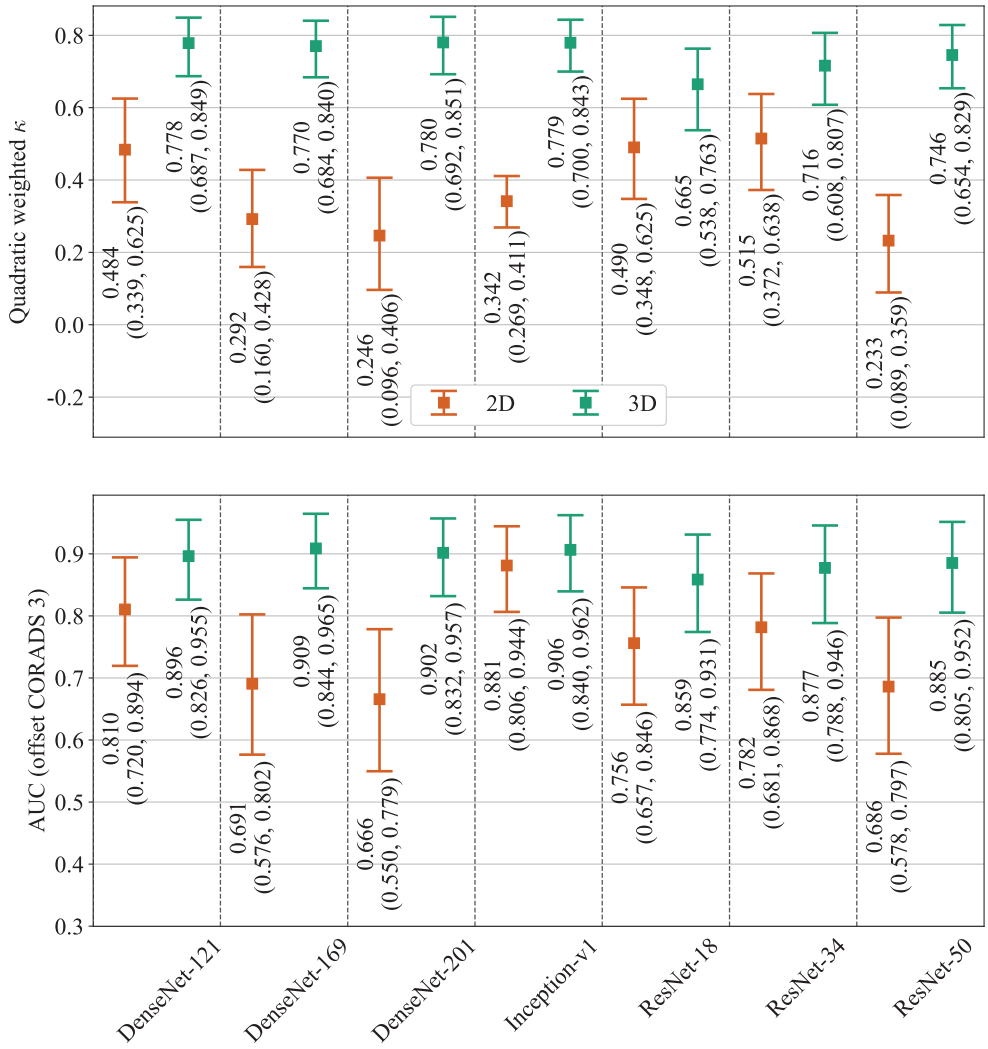


Figure 3.2: Performance of 2D and 3D CNN architectures on the internal test set for the task of CO-RADS grading from CT images is shown in QWK and AUC, respectively. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans).

3.4 Results

3.4.1 Architecture selection

Fig. 3.2 shows the QWK and AUC for the different 2D and 3D architectures. Table 3.3 shows the number of trainable parameters, single-model inference time for one sample and FLOP count for each architecture. All 2D architectures were outperformed by their 3D counterparts both in terms of QWK and AUC. The 3D DenseNet-201 architecture performed best in terms of QWK, followed by the 3D Inception-v1 architecture. In terms of AUC, the DenseNet-169 obtained the best performance, again followed by the 3D Inception-v1 architecture.

In the architecture selection, on average, training of the individual 3D models required approximately 26 700 iterations, while it required about 29 800 iterations for the 2D models.

Since the QWK takes into account the ordinal nature of the CO-RADS score, this metric

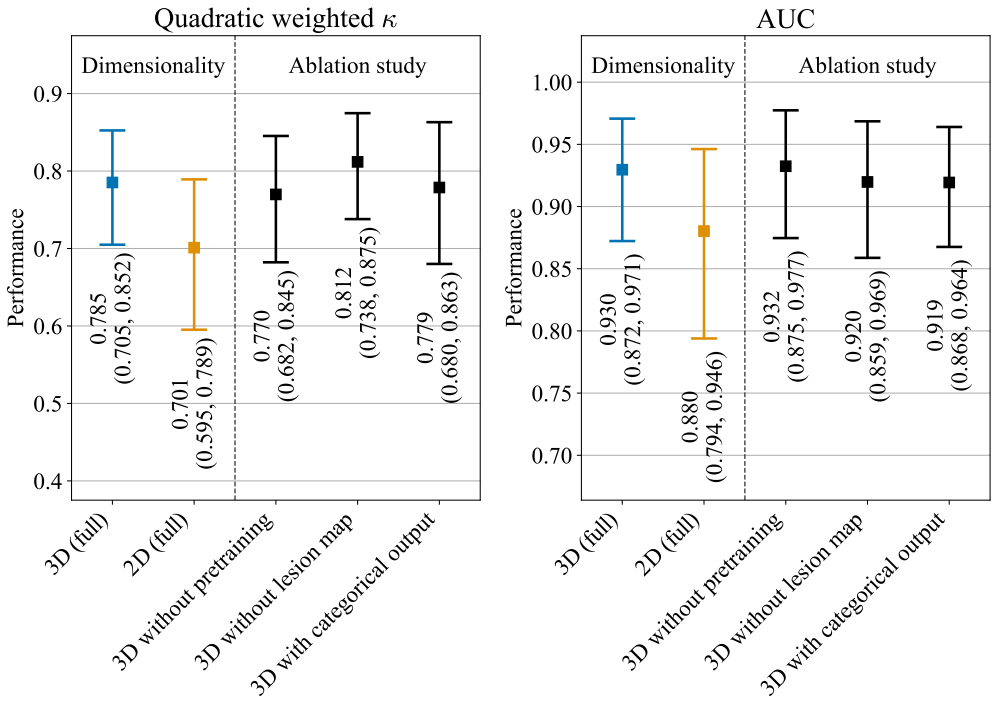


Figure 3.3: Comparison of 2D and 3D DenseNet-201 models and ablation study with this architecture for the task of CO-RADS grading from CT images. The analysis was performed on the internal test set. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans).

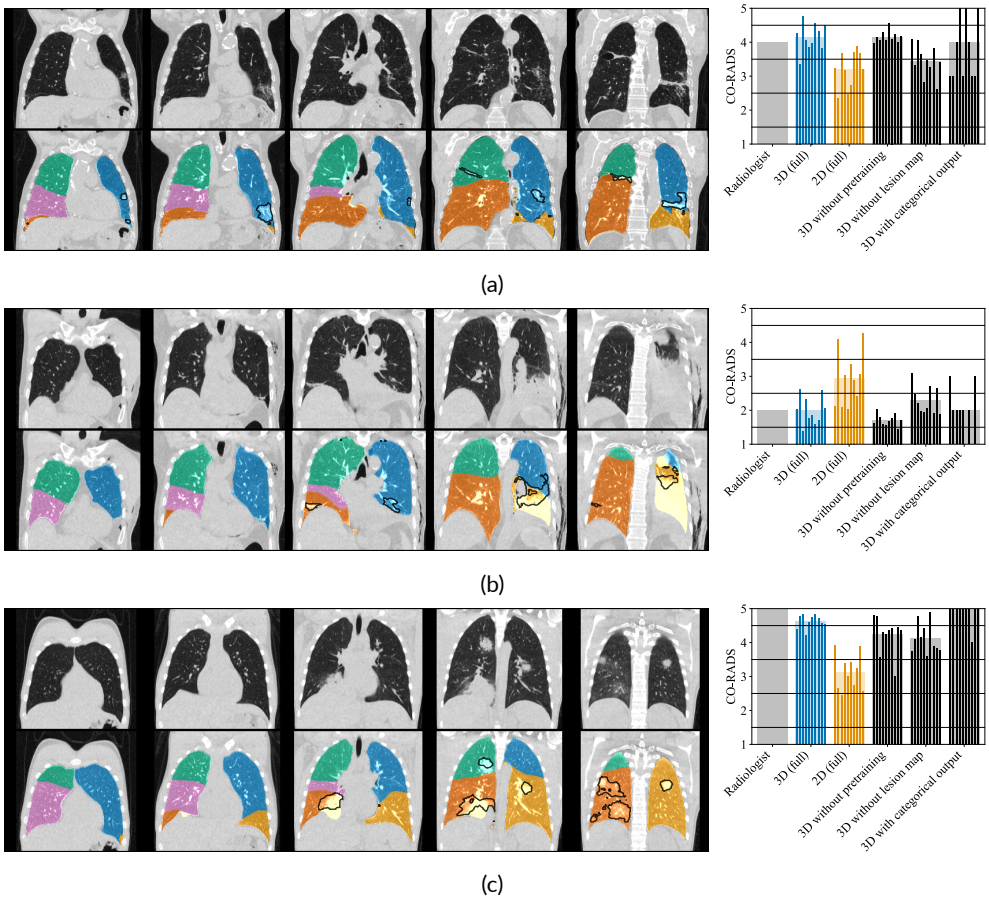


Figure 3.4: Example input-output pairs for the task of CO-RADS grading on the internal test set for the trained DenseNet-201 ensembles. Input examples are shown on the left. Top row: Coronal slices of an input CT scan. Bottom row: Lung segmentation used for centering and cropping are displayed with colored overlays. Delineations of the lesion masks that were used as a separate input channel are depicted as black lines. Output examples of the ensembles (wide, light bars) as well as the individual models these ensembles are composed of (narrow, dark bars) are shown on the right. **(a)** Radiology report: "GGO and consolidations especially lower lobes and posterior. Has had prior lung carcinoma. COVID-19 is probable, but other infection intrapulmonal is also possible." **(b)** Radiology report: "COVID-19 not probable, but also not ruled out. Known post-traumatic thorax, persistent pleura fluid, slice pneumothorax. Small amount of GGO and consolidation (left). Some pneumonia at thorax trauma, post-traumatic deviations." **(c)** Consolidation and GGO in all lobes. According to radiologist: "Very suggestive for COVID. Also positive PCR. Proven comorbidity."

was used to select the architecture to execute the ablation study with. In the rest of this section, we refer to the 3D DenseNet-201 ensemble as the 3D model and to the 2D DenseNet-201 ensemble as the 2D model.

3.4.2 2D vs. 3D CNNs

On the internal dataset, both the AUC and the QWK scores were significantly higher for the full 3D model (with transfer learning, lesion maps and continuous output) than for the full 2D model (p = 0.006 for AUC and p = 0.007 for QWK). Figures 3.3 and 3.6 show the corresponding CIs and ROC analyses respectively. Fig. 3.4 shows prediction examples from the full 3D, full 2D and ablated 3D models in blue, yellow, and black respectively.

We also trained an ensemble with the COVNet pipeline from Li *et al.* [32], which contains a ResNet-50 backbone that was pre-trained on ImageNet. With COVNet, we obtained a lower performance on the internal test set than when we applied the 3D model in our own pipeline. COVNet obtained a QWK of 0.567 (95% CI: 0.411-0.703, p = 0.004) and a lower AUC of 0.828 (95% CI: 0.741-0.906, p = 0.017) Our 2D model also outperformed COVNet in terms of both the QWK (p = 0.074) and AUC (p = 0.179).

Fig. 3.5 shows confusion matrices for the two dimensionalities. For 13 scans, the full 3D approach had predictions that were more than one CO-RADS category off. For the full 2D approach this was the case for 19 scans. Furthermore, the full 3D approach and 2D approach both had two cases that were further off than 2 categories.

3.4.3 Ablation study

The results of an ablation study to investigate the effect of each of the additional components added to the 3D CNN are shown in Fig. 3.3. The 3D model without ablations obtained an AUC of 0.930 (95% CI: 0.872-0.971) and a QWK of 0.785 (95% CI: 0.705-0.852). Removing any of the additions had a smaller effect on these performance metrics than changing the dimensionality of the architecture to 2D. Removing pre-training reduced the QWK to

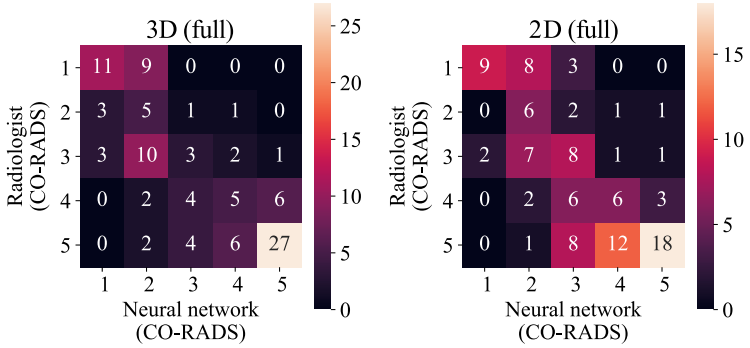


Figure 3.5: Confusion matrices for CO-RADS grading of the 2D and 3D DenseNet-201 model predictions on the internal test set. These models were trained with transfer learning, lesion maps and produced continuous output. The true label reference is from the radiology report. Cells contain the number of CT scans.

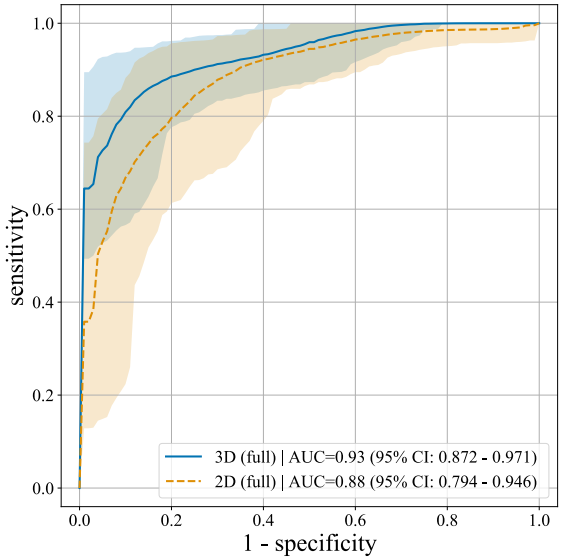


Figure 3.6: ROC analysis for the 2D and 3D DenseNet-201 models on the internal test set from Radboudumc (105 CT scans) for the task of CO-RADS grading. The analysis was performed with CO-RADS 1 and 2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans). It was performed for the full 2D and 3D models trained with transfer learning, lesion maps and continuous output.

0.770 (95% CI: 0.682-0.789, $p = 0.278$), but increased the AUC to 0.932 (95% CI: 0.857-0.977, $p = 0.428$). When the lesion segmentation input was removed from the model, the QWK was increased to 0.812 (95% CI: 0.738-0.875, $p = 0.091$) and the AUC was reduced to 0.920 (95% CI: 0.859-0.969, $p = 0.292$). Replacing the regression approach with a categorical target had a negative effect on both metrics, reducing the QWK to 0.799 (95% CI: 0.680-0.863, $p = 0.421$) and the AUC to 0.919 (95% CI: 0.868-0.964, $p = 0.324$). Fig. 3.4 shows prediction examples from the ablation study models in black.

The 3D model required 31 550 iterations for training on average. The 2D model, the network without pre-training, and the model without categorical output all required less iterations (25 650, 31 000 and 22 450, respectively). The model without lesion input required more iterations (32 750).

3.4.4 External evaluation

Fig. 3.7 shows the ROC curves of the full 3D and the full 2D model for the external iCTCF test set.

The 3D approach obtained an AUC of 0.919 (95% CI: 0.898-0.938) and outperformed the 2D approach that obtained an AUC of 0.915 (95% CI: 0.893-0.934, $p = 0.215$).

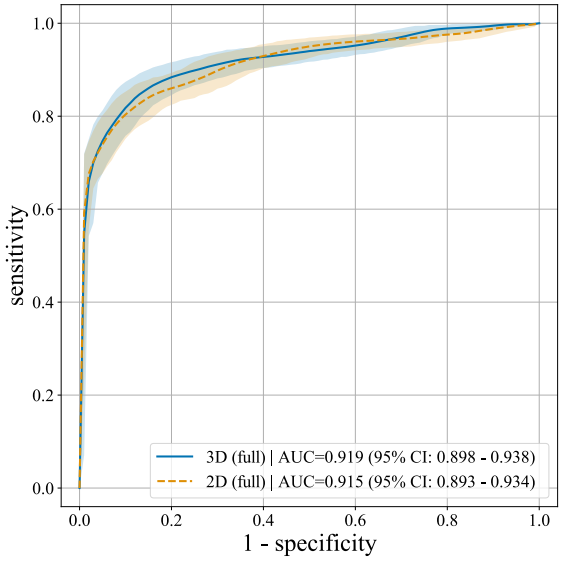


Figure 3.7: ROC analysis for the 2D and 3D DenseNet-201 models on the external iCTCF test set (742 CT scans) for the task of COVID-19 classification. The analysis was performed with 207 COVID-19 negative (Control) cases and 535 positive (Mild, Regular, Severe, Critically ill) cases.

3.4.5 Lesion segmentation model

For a single patch the lesion segmentation model inference time was 178.66 ms ± 14.56 ms, using 9.41×10^{11} FLOPs. The CT scans in the test set contained 12.8 patches on average. The model had 29.69×10^{6} parameters. Performance metrics for this model were reported by Lessmann *et al.* [40].

3.5 Discussion

In this paper, we identified and tested components of CNN based automated COVID-19 grading models. More specifically, we investigated how the performance of such models is affected by using different 2D and 3D CNN architectures, adopting pre-trained weights, using automatically computed lesion maps as additional network input, and predicting a continuous output instead of a categorical output. We evaluated all models with the same datasets to allow for a fair comparison between models.

Based on the architectures used in earlier automated COVID-19 classification research, we selected and compared the performance of the 2D and 3D variants of 7 CNN architectures for this task. We found that for all architecture types, the 2D models were outper-

formed by their 3D counterparts. The best performing model was a 3D DenseNet-201. In the rest of this section, we refer to the 3D DenseNet-201 as the 3D model and to the 2D DenseNet-201 as the 2D model.

The full 3D model (with transfer learning, lesion maps and continuous output) outperformed the full 2D model in terms of AUC and QWK score on the internal test set for COVID-19 classification and CO-RADS grading.

We compared our 2D model with COVNet, an architecture previously used in a similar COVID-19 classification task in CT [32], for which the authors reported an AUC of 0.96 for differentiating between COVID-19 positive and negative patients. The substantial difference between this result and our observations with COVNet illustrates the importance of using the same dataset when comparing different approaches.

We also observed a better diagnostic performance for COVID-19 classification by the 3D model on the external test set, although this performance increase was not statistically significant for a significance level of 0.05. AUC was 0.919 for the full 3D model, while it was 0.915 for the full 2D model. Ning *et al.* [41] developed a 2D model with slice-level annotations indicating if the slice was COVID-19 positive, negative or non-informative. Using a superset of the external set used in this paper for evaluation an AUC of 0.919 was obtained, which is the same as the AUC of our 3D model, even though our 3D model was trained with weaker labels and on data from a different population. This further emphasizes the importance of using 3D rather than 2D models.

The internal test set was comprised of data from the same population as the data the model was trained on, while the external test set was comprised of data from a different population. For the full 2D model, a lower AUC was obtained on the internal test set than on the external test set. This difference might be due to population differences between the internal and external test set, or due to the different definitions of the positive class, which were presence of COVID-19 and high suspicion of COVID-19 for the internal and external test sets respectively.

On the external test set, the full 3D model outperformed the full 2D model by a smaller margin in terms of AUC than on the internal dataset. This difference could be partly due to the different definitions of the positive class. However, we also found that it partly arises from the larger overall slice thickness in the external test set. All scans in the internal test set had a slice thickness of 0.5 mm. In contrast, 207 scans (40 COVID-19 positive, 167 negative scans) in the external test set had a slice thickness larger than 1.5 mm, which was the input resolution in our training and testing pipeline. When evaluating only on these scans, we obtained an AUC of 0.885 (95% CI: 0.835-0.931) for the full 3D model and an AUC of 0.891 (95% CI: 0.843-0.932) for the full 2D model. The external test set contained 535 scans (167

COVID-19 positive, 368 negative) with a slice thickness smaller than or equal to 1.5 mm. On these scans we obtained an AUC of 0.926 (95% CI: 0.902-0.947) for the full 3D model and an AUC of 0.918 (95% CI: 0.892-0.941) for the full 2D model. The performance of both models is lower for scans with a large slice thickness, but this effect is more apparent for the 3D model. Taking into account the increasingly smaller slice thickness of CT scans [109], this observation further supports our hypothesis that 3D models are better suited for COVID-19 grading applications than 2D models.

A possible explanation for why adding the extra dimension to the convolutions improves the performance is that it allows the CNN to take into account the 3D structure and full volume of individual lesions. This explanation is in line with the fact that radiologists typically use both the axial and coronal views to visualize the spread of COVID-19 related lesions across the lungs in CT scans, such as GGOs [29].

We could not directly compare the CO-RADS classification performance on the external set, since CO-RADS labels were not available. Moreover, the CO-RADS grading cannot be directly translated to the system used in the iCTCF dataset, since the former measures the probability of COVID-19 presence, while the latter quantifies the severity of the disease.

The ablation study on the internal test set showed that the further additions to the network and training procedure did not have a significant effect on the performance. Regardless of performance increases, using a continuous output removes the disadvantage of having to decide on the onset of the positive class for the predicted CO-RADS scores. Adding lesion maps as input and using inflated ImageNet weights for pre-training might both be ineffective for 3D automated CNN based COVID-19 grading methods.

The full 2D DenseNet-201 model obtained a better performance than the 2D DenseNet-201 model without pre-training, additional lesion map input, and continuous output. This indicates that some of these additional components positively affected the performance of the 2D model. However, even with all additional components, it was still outperformed by the vanilla 3D DenseNet-201.

We did not use clinical features available for the external dataset as input to the models trained in this work, since the main goal of this paper was to demonstrate the effect on performance of different COVID-19 grading and classification algorithm components.

3.6 Conclusion and future work

We compared a variety of 2D and 3D Convolutional Neural Network (CNN) architectures for COVID-19 classification from computed tomography scans and found that for all architectures considered, the 3D variants outperformed their 2D counterparts. We investigated

how the performances of the best performing architecture and its 2D counterpart were affected by including COVID-19 related lesion segmentations as additional input, using pretrained weights, and replacing the categorical output with a scalar continuous output.

We intentionally did not develop novel non-trivial architectural tweaks for small performance improvements, as many of them have been shown to be unnecessary and to not generalize well to other datasets and tasks [26, 27]. We leave systematic comparisons that explore other transfer learning schemes, make use of slice-level annotations, and use clinical features as model input for future work.

Radiologists can be aided in assessing CT scans on the presence of COVID-19 by automatic COVID-19 grading systems. This paper advances and speeds up the development of such systems in the following ways. Firstly, our findings aid in advancing the performance of automated COVID-19 grading systems and provide insight into the performance benefits of several of their components. These insights primarily indicate that future research and clinical applications should move towards using 3D CNNs for COVID-19 grading in CT scans. Secondly, the models and the automatic evaluation method used in this paper have been made available on the online Grand Challenge platform [111]. This allows researchers to obtain and compare the performance of their COVID-19 grading and classification solutions to other solutions on the platform. Thirdly, the output of all models used in this paper adheres to the standardized CO-RADS reporting system to facilitate easier integration into clinical workflow.

Acknowledgements

This publication is made possible in part by funding from the European Regional Development Fund (ERDF) East Netherlands.

PART II

Flagging ungradable and out-of-distribution data

"In the whole history of medicine there is no more beautiful episode than the invention of the ophthalmoscope, and physiology has few greater triumphs."

Edward Loring

4

International challenge on robust glaucoma screening

Authors: Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, RG Devika, PS Hrishikesh, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram Van Ginneken, Hans G. Lemij, Clara I. Sánchez

Original title: AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge

Published in: IEEE Transactions on Medical Imaging (Volume: 43, Issue: 1, Pages: 542-557, January 2024)

https://doi.org/10.1109/TMI.2023.3313786

Abstract

The early detection of glaucoma is essential in preventing visual impairment. Artificial intelligence (AI) can be used to analyze color fundus photographs (CFPs) in a cost-effective manner, making glaucoma screening more accessible. While AI models for glaucoma screening from CFPs have shown promising results in laboratory settings, their performance decreases significantly in real-world scenarios due to the presence of out-of-distribution and low-quality images. To address this issue, we propose the Artificial Intelligence for Robust Glaucoma Screening (AIROGS) challenge. This challenge includes a large dataset of around 113,000 images from about 60,000 patients and 500 different screening centers, and encourages the development of algorithms that are robust to ungradable and unexpected input data. We evaluated solutions from 14 teams in this paper and found that the best teams performed similarly to a set of 20 expert ophthalmologists and optometrists. The highest-scoring team achieved an area under the receiver operating characteristic curve of 0.99 (95% CI: 0.98-0.99) for detecting ungradable images on-the-fly. Additionally, many of the algorithms showed robust performance when tested on three other publicly available datasets. These results demonstrate the feasibility of robust AI-enabled glaucoma screening.

4.1 Introduction

GLAUCOMA IS ONE OF THE MAIN CAUSES of irreversible blindness and impaired vision in the world. It affects the optic nerve, which connects the eye with the brain, and leads to progressive visual field damage. This damage initially passes unnoticed by the patient. Only in later stages will glaucoma patients experience visual loss. According to estimates, by 2040, over 110 million people will have varying degrees of visual impairment caused by glaucoma [70], with 10% becoming blind in both eyes and 25% in one eye [129]. Many people experience visual impairment from glaucoma because it is often not detected until later stages [71, 72]. Current treatments of glaucoma cannot repair the damage, but can only halt or slow the progression of the condition [130]. Implementing screening programs to identify patients early on for treatment can alleviate the consequences of the disease. Artificial intelligence (AI) may be the enabling technology for the cost-effective implementation of these programs by automatically detecting perimetric glaucoma (*i.e.,* glaucoma in which there is already visual field damage) in color fundus photographs (CFPs) [131–135].

Existing AI solutions have been shown to drop in performance in real-world screening practice due to comorbidities, poor quality images, different ethnicities, or unexpected out-of-distribution (OOD) samples [134]. Ad-hoc quality check modules have been added to AI solutions to overcome this performance drop, but recent research has indicated that

Table 4.1: Statistics of the Rotterdam EyePACS AIROGS dataset. # = number of. CFPs = color fundus photographs.

		Full set	Training set	Test set
# CFPs / # patients		112,732 / 60,071	101,442 / 54,274	11,290 / 5,797
Prevalence (# CFPs (% within set) / # patients)	RG	4,872 (4.3%) / 3,531	3,270 (3.2%) / 2,336	1,602 (14.2%) / 1,195
	NRG	106,306 (94.3%) / 58,388	98,172 (96.8%) / 53,400	8,134 (72.0%) / 4,988
	U	1,554 (1.4%) / 1,415	0 (0.0%) / 0	1,554 (13.8%) / 1,415
Age at encounter (mean $\pm$ std. dev.)	All classes	56.9 $\pm$ 10.3	56.7 $\pm$ 10.2	59.3 $\pm$ 10.9
	RG	63.3 $\pm$ 10.5	63.0 $\pm$ 10.5	63.9 $\pm$ 10.7
	NRG	56.5 $\pm$ 10.2	56.5 $\pm$ 10.1	57.5 $\pm$ 10.4
	U	63.7 $\pm$ 11.0	N/A	63.7 $\pm$ 11.0
# sites		500	486	376
Cameras (# CFPs (% within set) / # patients)	Canon CR1	11,462 (10.2%) / 6,013	10,274 (10.1%) / 5,422	1,188 (10.5%) / 591
	Canon CR2	10,523 (9.3%) / 5,538	9,179 (9.0%) / 4,866	1,344 (11.9%) / 672
	Canon DGI	10,644 (9.4%) / 5,690	9,581 (9.4%) / 5,145	1,063 (9.4%) / 545
	Optovue iCam 100	29,108 (25.8%) / 16,166	26,480 (26.1%) / 14,742	2,628 (23.3%) / 1,424
	TopCon NW200	3,109 (2.8%) / 1,588	2,888 (2.8%) / 1,478	221 (2.0%) / 110
	TopCon NW400	22,519 (20.0%) / 11,736	20,557 (20.3%) / 10,737	1,962 (17.4%) / 999
	Centervue DRS	1,805 (1.6%) / 988	1,598 (1.6%) / 879	207 (1.8%) / 109
	Nidek AFC300	61 (0.1%) / 31	53 (0.1%) / 27	8 (0.1%) / 4
	Crystalvue NFC-700	8 (0.0%) / 4	6 (0.0%) / 3	2 (0.0%) / 1
	Unknown	23,493 (20.8%) / 12,323	20,826 (20.5%) / 10,981	2,667 (23.6%) / 1,342

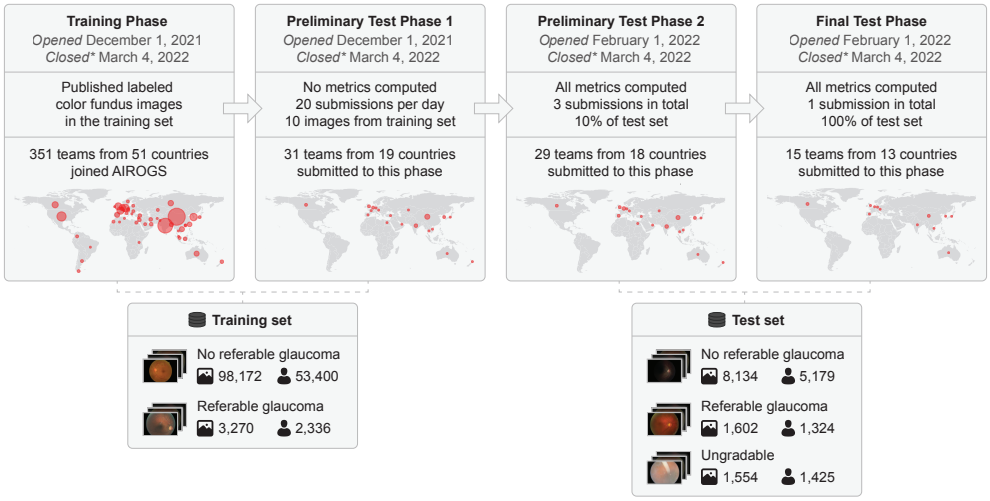


Figure 4.1: Overview of all phases in the AIROGS challenge. A world map is shown for each phase that indicates with red circles from which countries the teams that participated in that phase originated. A circle is shown for each country from which at least one team participated and its size represents the number of teams that joined from that country. The relevant subset of the AIROGS dataset for each phase is shown at the bottom of the figure. *All phases reopened for new submissions after the winning teams were announced.

these quality checks are not sufficiently accurate when deployed in real-world settings [23]. To allow a safe and effective deployment in screening, the reliability and robustness of such solutions need to be assessed. Medical image analysis challenges often exclusively focuses on performance metrics that are potentially unrealistic and overestimated due to the use of test sets that do not represent real-world scenarios. Moreover, metrics to measure reliability and robustness are often neglected due to the difficulty of estimating them in the provided test sets.

To develop solutions that overcome the aforementioned issues related to robustness in glaucoma screening, we organized the Artificial Intelligence for RObust Glaucoma Screening (AIROGS) challenge. The goal of this challenge was to evaluate the feasibility of the development of a state-of-the-art, reliable AI solution that takes a CFP as input and provides as output the likelihood of referable glaucoma, accompanied with outputs for robustness (*i.e.,* predicting whether the input image can be graded reliably or not). The screening task was to distinguish no referable glaucoma (*i.e.,* either no glaucoma at all or non-referable glaucoma, that is suspected pre-perimetric glaucoma) from referable glaucoma. This is different from, for example, distinguishing between multiple glaucoma severity stages.

To encourage the development of solutions that are robust to any kind of ungradable and unexpected input data and are equipped with inherent robustness mechanisms, the training set we provided was a subset of the full AIROGS dataset where only gradable im-

ages were included and ungradable images excluded. The test set, however, is unfiltered, containing all images found in screening settings (gradable and ungradable), representing a real-world scenario.

AIROGS was part of the International Symposium on Biomedical Imaging (ISBI) 2022 challenge program. It reopened after presenting the results during ISBI 2022 and submissions can still be made on Grand Challenge*.

Our challenge, along with the dataset we made publicly available, distinguishes itself from previous glaucoma challenges. First, to the best of our knowledge, our dataset is the largest publicly available CFP dataset with glaucoma labels by a large margin. In total, our dataset contains 112,732 CFPs, exceeding the size of other publicly available datasets containing CFPs with glaucoma, of which the sizes range from 22 to 2,000 CFPs [136–144]. It is a highly diverse dataset as it originates from 500 screening centers across the United States of America and was acquired with a large variety of cameras. Second, the AIROGS challenge is the first challenge to emphasize robustness in glaucoma screening. Third, AIROGS is one of the first types of challenges on grand-challenge.org that requires participants to submit an algorithm (a *Type 2* challenge), rather than a file with their predictions on the test set (a *Type 1* challenge), as is done in more traditional challenges. This makes human intervention in the generation of test set results impossible, reducing the possibility of cheating. Moreover, it greatly improves reproducibility, allowing everyone to reuse the trained algorithms that were submitted and apply them to new data in a cloud-based environment. Fourth, the reproducibility enabled testing of the participating algorithms on three external datasets: two for evaluating the screening task and one for evaluating robustness.

4.2 Datasets

4.2.1 The Rotterdam EyePACS AIROGS dataset

The Rotterdam EyePACS AIROGS dataset contains 112,732 CFPs from 60,071 subjects and 500 different sites with a heterogeneous ethnicity. The images were originally acquired for a diabetic retinopathy screening program [145]. For grading of the CFPs, all graders were trained and then selected for this task using the European Optic Disc Assessment Trial (EODAT) [146], containing 110 stereoscopic optic nerve photographs, in which all glaucomatous eyes had reproducible visual field defects on standard automated perimetry. 90 experienced ophthalmologists and optometrists were examined and those who scored at

*https://airogs.grand-challenge.org

least 85% overall accuracy and 92% specificity were selected to label images for the present study. Eventually, 30 out of 90 candidates passed.

For each eye, three images were taken by the camera operators to reduce the number of ungradable eyes. When labeling the images, graders classified one eye at a time. The labeling tool first presented the first CFP for each eye, upon when graders could choose from the options "Referable glaucoma" (RG), "No referable glaucoma" (NRG), or "Ungradable" (U). If a grader selected U for the first image, the tool showed the consecutive CFP. The third image was presented if the second image was also deemed U. Each eye was scored by two separate graders, who were both unaware of the identity of the other grader. If the two graders agreed on the label of a CFP, this became the final label. If they disagreed, the image was scored by one of the glaucoma specialists who passed the EODAT test with at least 95% accuracy. The final label was then based on his judgment.

The graders were instructed to select RG if they found glaucomatous signs which they expected to be associated with visual field defects on standard automated perimetry. The signs that could be selected were "appearance neuroretinal rim superiorly", "appearance neuroretinal rim inferiorly", "baring of the circumlinear vessel superiorly", "baring of the circumlinear vessel inferiorly", "disc hemorrhage(s)", "retinal nerve fiber layer defect superiorly", "retinal nerve fiber layer defect inferiorly", "nasalization (nasal displacement) of the vessel trunk", "laminar dots" and "large cup". If the graders did not expect any glaucomatous visual field defects, NRG was to be selected, ignoring any comorbidities (*e.g.,* age-related macular degeneration and diabetic retinopathy). If there was not enough information visible in the CFP to decide between RG and NRG, graders were instructed to select U. Since the goal of this study was to develop solutions for automated screening, glaucoma severity was not reported by the human graders.

The graders were not only evaluated at the start, but they were periodically monitored during the grading process, as well. If their sensitivity or specificity dropped below 80% or 95%, respectively, they were removed from the study and all images they labeled were re-graded by any of the remaining graders. In case a grader wrongly classified a CFP as U, while its final label was NRG or RG, their specificity or sensitivity went down, respectively. In the end, 20 graders remained.

Out of the three CFPs that were available for each eye, we only included the RG or NRG photograph in the dataset if it was available. Otherwise, only one of the U photographs was used. We split the data into a training set of 101,442 CFPs and a test set of 11,290 CFPs, ensuring that data from patients in the training set was not in the test set. We randomly sampled patients when making the split, oversampling patients with ungradable and RG CFPs for the test set, such that approximately 1,600 RG and 1,600 U photographs ended up

in the test set. When making the split, we ensured that data from a single patient ended up in either the training set or the test set. Since we were interested in AI solutions that can identify ungradable data without training on ungradable data, we left out all U photographs that ended up in the AIROGS training set. Table 4.1 shows statistics about RG, NRG, and U prevalence, age, sites, and cameras for the full dataset, the training set, and the test set. For further information about the acquisition process, labeling, and dataset statistics, including the prevalence of ethnicity, please refer to the paper on the REGAIS dataset, of which AIROGS is a subset [147]. The AIROGS dataset includes 99% of the data from the REGAIS dataset. The difference in size is due to the exclusion of ungradable eyes of patients in the AIROGS set, which was not done in the REGAIS set. The dataset included data from "people of African descent, Whites, Asians, Latin Americans, native Americans, people from the Indian subcontinent, people of mixed ethnicity, and people of unspecified ethnicity" [147]. Approval from the Institutional Review Board of the Rotterdam Eye Hospital was obtained to conduct this research.

4.2.2 External datasets

The participants uploaded their trained algorithms, rather than a file with predictions on our test set, to our challenge platform. This enabled us to reuse the developed models on external data after the challenge ended. To evaluate model generalization and to demonstrate this reusability, we applied all trained algorithms to three external datasets: Retinal Fundus Glaucoma Challenge (REFUGE) [142], Glaucoma grAding from Multi-Modality imAges (GAMMA) [144] and Diabetic Retinopathy Image Database (DRIMDB) [148]. The former two are datasets with positive and negative glaucoma CFPs, which we used for externally evaluating the screening performance. We used the latter dataset, which contained different types of ungradable images, to evaluate the robustness externally.

The REFUGE test set contained 400 CFPs, of which 40 CFPs showed glaucoma and 360 CFPs did not. The definition of glaucoma was glaucomatous damage in the optic nerve head area and reproducible glaucomatous visual field defects, which is similar to our definition of glaucoma described earlier [142].

The GAMMA dataset is a multi-modal dataset with optical coherence tomography scans and CFPs for each eye. We used the CFP data from the 100-sample training set as only that subset of the GAMMA dataset had publicly available labels. We defined positive glaucoma in the same way as Wu *et al.* [144], *i.e.*, as the union of the early, intermediate, and advanced glaucoma stages. These stages were defined using the mean deviation (MD) from the visual field reports as follows: an MD less than -6 dB the early stage, an MD between -6 dB and -12 dB for the intermediate stage, and an MD worse than -12 dB for the advanced stage [144].

This resulted in 50 negative and 50 positive glaucoma samples.

DRIMDB is a dataset with 125 "Good" CFPs, 69 "Bad" CFPs, and 22 "Outlier" CFPs. According to Şevik *et al.* [148], one of the criteria of the "Good" category was OD presence. We also manually confirmed the OD was visible in all CFPs that were labeled "Good" in the DRIMDB dataset. Therefore, we assumed the CFPs with the category "Good" were gradable. The images labeled "Bad" and "Outlier" were assumed to be ungradable. This resulted in 125 gradable and 91 ungradable images.

4.3 Challenge setup

The AIROGS challenge consisted of four phases (see Fig. 4.1). The *Training Phase* opened on the 1st of December 2021 and closed on the 4th of March 2022, providing the participants with approximately three months to develop their solutions. At the start of this phase, the training set was released and has since been available for download under the CC BY-NC-ND license on Zenodo[†].

To ensure fair competition and to encourage the development of inherent robustness mechanisms, teams were not permitted to use additional fundus image training data, including weights pre-trained on fundus image data or in pre-processing steps such as OD segmentation. Manually labeling the challenge data and using the resulting annotations during training was allowed.

To test the algorithms developed by participants, they needed to wrap their trained algorithm in a Docker[‡] container and submit it to our challenge platform. This allows the submitted algorithms to be run on data that is not directly accessible by the participating teams. Example code for generating such a containerized submission can be found on GitHub[§]. *Preliminary Test Phase 1* opened and closed simultaneously with the *Training phase* and served as a check for whether the submitted algorithms could be run on the challenge platform and produced the output in the expected format. Algorithms were tested on 10 images from the training set for this check. All algorithms were executed on the challenge platform using an NVIDIA T4 GPU (16 GB VRAM) with 8 CPUs (32 GB RAM).

The test set was and still is closed, meaning the image data and the labels are private and cannot be downloaded. *Preliminary Test Phase 2* opened on the 1st of February 2022 and we allowed three submissions per team to this phase, as it used 10% of the test set for evaluation. All challenge metrics were also computed and reported back to the participants. The *Final Test Phase* opened simultaneously with *Preliminary Test Phase 2*, but algorithms

[†] https://zenodo.org/record/5793241
[‡] https://docker.com
[§] https://github.com/qurAI-amsterdam/airogs-example-algorithm

were tested on 100% of the test data and only one submission per team was allowed. The challenge metrics computed for this phase were used for the final team ranking.

The algorithms were expected to produce four outputs, of which two were related to glaucoma screening performance (*i.e.,* image classification of RG and NRG) and the other two to robustness (*i.e.,* the identification of U). The glaucoma screening outputs were a likelihood score for RG (O_1) and a binary decision for RG (O_2, positive if RG and negative if NRG). The ungradability outputs were a binary decision on whether the image is ungradable (O_3, positive if ungradable and negative if ungradable) and a non-thresholded scalar value that is positively correlated with the likelihood for ungradability (*e.g.,* the entropy of a probability vector produced by a machine learning model or the variance of an ensemble) (O_4). Output O_2 was not used in the evaluation pipeline for the challenge leaderboard, but it was requested by the challenge organizers for further analysis.

The evaluation was also based on the two aspects of screening performance and robustness, with two metrics per aspect. Screening performance was evaluated using the standardized partial area under the receiver operating characteristic curve [149] (90-100% specificity) for RG ($pAUC_S$), and the sensitivity at 95% specificity ($SE@95SP_S$). These metrics were based on these specificity ranges, as a high specificity is required for cost-effective glaucoma screening due to its relatively low prevalence [150, 151]. $pAUC_S$ and $SE@95SP_S$ are both based on output O_1. For evaluating the robustness, we determined the model's agreement with the human reference on ungradability using Cohen's kappa score (κ_U), calculated using output O_3. Furthermore, we calculated the area under the receiver operator characteristic curve using the human reference for ungradability as the true labels and output O_4 as the target scores (AUC_U).

To determine the final ranking, we first ranked all participants on the four individual metrics $pAUC_S$, $SE@95SP_S$, κ_U, and AUC_U resulting in the rankings R_{pAUC_S}, $R_{SE@95SP_S}$, R_{κ_U}, and R_{AUC_U}, respectively. The final score S_{final} was then calculated as the mean of those rankings:

$$S_{final} = \frac{R_{pAUC_S} + R_{SE@95SP_S} + R_{\kappa_U} + R_{AUC_U}}{4}. \tag{4.1}$$

The final ranking (later also referred to as *Mean position*), was based on S_{final}, where a lower value for S_{final} resulted in a higher ranking.

We calculated 95% confidence intervals (CIs) with non-parametric bootstrapping using 1000 iterations [152]. The code for evaluating submissions can be found on GitHub[¶]. The performance of human graders was calculated by comparing the labels given by the individual graders (excluding the two glaucoma specialists, since the final labels were equal to

[¶]https://github.com/qurAI-amsterdam/airogs-evaluation

their decision in case of disagreement) to the final labels as defined in Section 4.2.1. To compute the performance of all human graders combined, each image was weighted equally in the calculation of the metrics. We also evaluated ensembles of participating algorithms, which were generated by averaging the outputs of these algorithms.

4.4 Participating methods

Fifteen teams submitted a working solution to the *Final Test Phase*, of which one team did not opt-in to contribute to the current paper. In this section, we present the methods of the fourteen participating teams. More extensive descriptions are available on the AIROGS challenge website**. Table 4.2 and 4.3 summarize the participating methods in a structured manner.

4.4.1 PUMCH-eye [161]

The *PUMCH-eye* team proposed an approach with five trained models in their workflow. The first model (M_{disc}) was a segmentation model with ResNet101-UperNet [162] as the backbone that segmented the OD in the input CFP. For the development of this model, they manually labeled the OD in 40 images. In case M_{disc} successfully detected the OD, they computed the center c and the diameter d of the segmentation to crop the input image around c with size $3d$. This cropped image was then fed into a vision transformer [163] for the binary classification of RG and NRG. If the OD detection was unsuccessful, they fed the original input image to a different vision transformer for binary classification of RG and NRG.

The team also developed a vessel segmentation model with 40 images in which they manually annotated vessels (M_{vessel}). They trained a ResNet-18 (R_{vessel}) which took the output of M_{vessel} as input data, using the first 500 images in the training dataset and 100 manually selected images in the training set with relatively poor image quality. This classification model served as one of the inputs for ungradability classification. The second input was taken from M_{disc}. The ungradability likelihood output (O_4) was then defined as the output likelihood of the binary classification model R_{vessel} (*i.e.,* O_{vessel}) or, if M_{disc} could not detect an OD, as $R_{vessel} + 0.75$. O_3 was positive if O_4 was at least 0.95 and negative otherwise. The vessel segmentation model was evaluated on four randomly selected images, on which a Dice score of 0.787 was achieved. This was lower than the state-of-the-art for this problem [164]. However, this was expected to be sufficient for the downstream task of ungradability detection.

**https://airogs.grand-challenge.org/evaluation/final-test-phase/leaderboard/

Table 4.2: Method overview from all participating teams for the screening task. OD = optic disc. #ODs = number of CFPs in which the OD was manually labeled. #vessels = number of CFPs in which the vessels were manually labeled.

		Screening																																																
		Architecture																Pre-processing			Manual labels			Loss function			Optimizer				Pre-training			Class imbalance solution		Data augmentation during training														
#	Team	Vision transformer	Swin-Transformer-B	EfficientNet	DeiT-S	MobileNet	ResNet	ResNet-RS	SeResNeXt	VGG	DenseNet	Inception-V3	ConvNeXt	SeNet	SeResNet	Inception-ResNet-v2	Ensemble	Crop field-of-view	Crop OD	Histogram equalization	#ODs	#vessels	Input size of classification model	Cross-entropy	Focal loss	Kullback-Leibler	Adam [153]	AdamW [154]	SGD	SAM [155]	ImageNet	Masked auto-encoder [156]	Noisy student [157]	Balanced sampling	Weighted loss	Brightness modifications	Color jitter	Elastic deformations	Flipping	Blurring	Rotation	Translation	Zooming	Shearing	Quality reduction	Sharpening	Erasing	AutoAugment [158]	RandAugment [159]	TrivialAugment [160]
1	PUMCH-eye	✓																	✓		40	40	384^2	✓				✓				✓																✓		
2	RWTH-CuP		✓	✓	✓												✓		✓	✓	3,221		224^2	✓			✓	✓			✓			✓				✓	✓						✓					
2	Eyelab	✓																	✓		1,500		384^2	✓					✓					✓		✓			✓		✓	✓	✓	✓			✓			
4	Tien			✓							✓						✓	✓					$1{,}024^2$	✓			✓							✓					✓		✓		✓							
5	UPF+AIML					✓											✓						512^2	✓				✓		✓	✓																			✓
6	FMS-CETCV						✓																512^2		✓				✓		✓								✓											
7	ICT_HCI						✓											✓					512^2	✓			✓				✓			✓															✓	
8	SK							✓									✓	✓					256^2	✓			✓				✓				✓		✓			✓	✓	✓	✓							
9	SACM								✓	✓	✓	✓					✓		✓		735		$120^2/224^2$	✓				✓	✓		✓				✓															
10	UPRetina-UR			✓				✓										✓					512^2		✓		✓				✓			✓															✓	
11	OPTIMATeam											✓											224^2			✓	✓				✓			✓		✓			✓	✓	✓	✓	✓							
11	MA			✓					✓		✓			✓	✓	✓		✓					512^2	✓			✓				✓			✓		✓			✓	✓	✓	✓	✓			✓				
13	YC										✓						✓		✓	✓	101		608^2	✓					✓				✓		✓	✓			✓	✓	✓	✓	✓							
14	Mirazzak	✓											✓				✓						$224^2/384^2$	✓			✓				✓				✓				✓		✓									

Table 4.3: Method overview from all participating teams for the ungradability task and the deep learning frameworks they used. OD = optic disc. AE = autoencoder. VAE = variational autoencoder. rec. error = reconstruction error. OOD = out-of-distribution.

		Robustness				General
#	Team	Method	OD detection for uncertainty or confidence estimation	Threshold based on manual identification of low-quality images in the development set	Input size for ungradability approach	Deep learning framework
1	PUMCH-eye	Vessel and OD segmentation	✓	✓	384^2	PyTorch
2	RWTH-CuP	OD detection confidence	✓	✓	224^2	PyTorch
2	Eyelab	OD presence detection	✓	✓	384^2	PyTorch
4	Tien	AE rec. error + probability classification model			$1{,}024^2$	PyTorch
5	UPF+AIML	Synthetic image degradations		✓	512^2	PyTorch
6	FMS-CETCV	Interpolated Gaussian descriptor			256^2	PyTorch
7	ICT_HCI	Probability classification model		✓	512^2	PyTorch
8	SK	Energy-based OOD + activation rectification			256^2	PyTorch+Lightning
9	SACM	AE and VAE rec. error + OD detection confidence	✓	✓	288^2	PyTorch
10	UPRetina-UR	Test-time augmentation + probability classification model			512^2	PyTorch+fast.ai
11	OPTIMATeam	Deep Dirichlet uncertainty			224^2	TensorFlow+Keras
11	MA	Ensemble			512^2	TensorFlow+Keras
13	YC	Monte-Carlo Dropout			608^2	TensorFlow
14	Mirazzak	Regret function			224^2/384^2	PyTorch

4.4.2 RWTH-CuP [165]

The *RWTH-CuP* team proposed an approach with two steps consisting of cropping around the OD by employing a detection network, followed by an ensemble of transformers (Swin Transformer-B [166] and DeiT-S [167]) and convolutional neural networks (CNNs) (EfficientNet-B4 [118] and EfficientNetV2-M [168]) that classifies the cropped image. They manually labeled the OD and its environment in 3,221 CFPs to develop this detection network, for which they trained a YOLOv5 [169] object detector network.

For ungradability classification, the team used a hybrid approach. As the probability that an image is ungradable is high if the OD could be found by the object detector network, they employed the confidence score of the YOLOv5 detection model as one of the ungradability measures. To capture other ungradability causes, such as blurred depictions of the OD, they trained an additional classifier on a manually selected subset of the CFPs in the development set. The team considered the 4000 CFPs with the lowest confidence score of the object detector and manually selected 600 images that were assumed by the team to be very close to being classified as ungradable. They used another set of 2,000 high-quality images to train an EfficientNet-B4 [118] ungradability classification model. O_4 was then defined as $(1 - c) + g$, where c is the object detection confidence and g the output the ungradability classification model. The binary O_3 output value was determined using a cut-off manually determined by a medical doctor in 20,000 images from the development set for which O_4 was computed.

4.4.3 Eyelab [170]

The *Eyelab* team employed a two-stage approach for glaucoma classification. The first step was to detect and crop the OD area and the second step was a vision transformer [171] that classified the cropped image from the first step. For the detection model, they trained a YOLOv5 [169] model using semi-automatically generated labels. Their method for ungradability detection was based on whether the optic disc detection model from the first step found an optic disc to be present.

4.4.4 Tien [172]

Tien used an ensemble of an EfficientNet [118] and DenseNet [115] for the classification of RG and NRG. For the ungradability task, they used an autoencoder network and a blending engine. They used the reconstruction error as a measure of the likelihood of ungradability. The higher the reconstruction error, the more likely it is that the image is ungradable. The blending engine fused the probability output from the binary classification model as

a weight factor to the reconstruction error. The highest weight was 1 (when the probability was 0.5) and the weight was lowest when the probability is certain (either 0 or 1).

4.4.5 UPF+AIML [173]

Team *UPF+AIML* trained two separate models, both based on the MobileNet-V2 [174] architecture for lightweight training, and both optimized with the Sharpness-Aware Minimization (SAM) [155] technique for better generalization. The first model was trained on the available training set for the screening task. The second model was tasked with identifying out-of-distribution data, *i.e.,* ungradable images in this case. For this, ungradable images were simulated by applying four image transformations (brightness, gamma, saturation, and blur) online to the data, with such a strength that they would destroy image content and turn images useless for diagnostic purposes. The ungradability detection model was trained on a mixture of gradable (sampled directly from the original training set) and ungradable (simulated). After training, this model was applied to the training set, where all images were expected to be gradable. The threshold that would classify 0.1% of the training set as ungradable was selected for ungradability detection.

4.4.6 FMS-CETCV [175]

The *FMS-CETCV* team used a binary classifier with ResNet-50 as the backbone for classifying RG and NRG. They used focal loss [176] to account for the class imbalance in the training set.

For the classification of ungradable images, they used a self-supervised learning approach, inspired by the work of Oza *et al.* [177], where a one-class classification method was presented for unsupervised anomaly detection. The one-class classifier builds a feature space by extracting the features of the training sample which contain only the positive samples (*i.e.,* gradable images). They used an encoder with ResNet-18 as the backbone, which is trained on the AIROGS test set that only contains gradable images. The feature space produced by this encoder is then used by a Gaussian anomaly classifier to distinguish gradable and ungradable images.

4.4.7 ICT_HCI [178]

Team *ICT_HCI* used ResNet-50 for their RG and NRG classification model. During inference, they made five random 512×512 crops of the image and then provided all crops separately to the model and get five scores. If the maximum of five scores was greater than 0.9,

they let it be the output of the model, otherwise, they took the mean of the five scores as the output of the model.

The team used the minimum class probability of the two classes RG and NRG as the likelihood for ungradability. If and only if the ungradability likelihood was greater than 0.1, they set O_3 to be positive.

4.4.8 SK [179]

The *SK* team employed ResNet-RS [180] for RG and NRG classification. They replaced the final linear layer of ResNet-RS with a single linear layer with two channel outputs.

For ungradability classification, they used an inference-time OOD energy-based method [181] combined with activation rectification [182]. The energy-based method uses a scoring function based on energy, instead of softmax, to discriminate in-distribution (ID) and OOD data. In activation rectification, the outsized activation of a few layers can be attenuated by rectifying the activations at an upper limit. After rectification, the output distributions for ID and OOD data become much more well-separated. It is based on the observation that the mean activation for ID data is well-behaved with a near-constant mean and standard deviation, and the mean activation for OOD data has significantly larger variations across units and is biased towards having sharp positive values.

4.4.9 SACM [183]

Team *SACM* used the YOLOv5 [169] detection model to crop the optic disc in the CFP input image. In a semi-automated process, they manually labeled the locations of 735 optic discs and trained the detection model with 4,088 in total. The cropped image was then passed through an ensemble of classifiers (SeResNext-50 [184], VGG-16 [6], DenseNet-161 [115], EfficientNet-B5 [118], EfficientNet-B7 [118] and Inception-V3 [185]) to make the final prediction. They also used test-time augmentation.

For the robustness task, they used the detection model confidence, an autoencoder, and a variational autoencoder (VAE) [186]. They combined these three aspects of their pipeline using this formula to achieve a final ungradability score for O_4 as $(1 - c) \cdot s \cdot p_{autoencoder} \cdot p_{vae}$, where c refers to the detection network confidence, and $p_{autoencoder}$ and p_{vae} refer to the mean squared error between the input and output of the autoencoder and VAE, respectively.

4.4.10 UPRetina-UR [187]

The *UPRetina-UR* team used ResNet-RS-50 [180] for the classification of RG and NRG. They oversampled cases with RG during training to account for the class imbalance.

They employed a closed-set classification approach for the ungradability task based on the method proposed by Vaze *et al.* [188]. They applied test-time augmentation to obtain five predictions that are averaged to produce O_4.

4.4.11 OPTIMATeam [189]

OPTIMATeam used the first two blocks from the Inception-V3 [185] network for the classification of RG and NRG. They only used these two blocks to reduce the receptive field size, which was necessary for their ungradability approach.

The ungradability approach was based on the direct modeling of the uncertainty following the evidential deep learning approach [190]. They used Deep Dirichlet uncertainty estimation as the ungradability score O_4. To set a threshold for getting a binary value for O_3 based on O_4, they assumed that diagnosis is only possible if the OD has enough image quality for diagnosis, as glaucoma's main structural manifestation occurs in that region. They applied Grad-CAM [191] on the trained model for the screening task and occluded out the region where Grad-CAM was greater than 0.5. This allowed them to produce ID and OOD samples in their validation set, with which they computed the threshold for the binary ungradability decision. In particular, they constructed a receiver operating characteristic (ROC) curve using their values for O_4 with these ID and OOD samples. The ROC threshold where the sensitivity was 0.5 was set to calculate O_3.

4.4.12 MA [192]

Team *MA* used an ensemble of these twelve different architectures for the glaucoma screening task: SeNet-154 [184], SeResNet-101 [184], SeResNeXt-101 [184], EfficientNet-B1 [118], EfficientNet-B2 [118], EfficientNet-B3 [118], EfficientNet-B4 [118], EfficientNet-B5 [118], EfficientNet-B6 [118], EfficientNet-B7 [118], DenseNet-201 [115], Inception-ResNet-v2 [116]. The RG likelihood was computed by averaging the likelihoods of all respective models in the ensemble.

The ungradability output O_4 was the sum of the variances between all models in the ensemble for the positive and negative class probabilities. O_3 was positive if O_4 exceeded 0.2 and negative otherwise.

4.4.13 YC [193]

The *YC* team used two DenseNet-121 networks to classify RG and NRG in the CFPs. The first network was trained with the full CFP as input and the second network used a version of the CFP that was cropped around the optic disc as input. After the last convolutions of these networks, a fully connected layer with dropout was added. The outputs of these fully connected layers were then concatenated and used as the input to another fully connected layer with dropout, which was followed by the final layer of the network. For cropping the CFPs around the optic disc, they trained a U-Net [9] with a DenseNet-121 [115] backbone. To train this segmentation network, they first roughly annotated the position of the optic disc in 101 CFPs in the training set. Subsequently, they generated reference segmentation maps using a probability density function of the multivariate normal distribution around the annotated optic disc position.

They used Monte-Carlo drop-out [194] with 20 predicted probabilities per image for the robustness task. Then they statistically tested a Wilcoxon one-sample test whether the mean of the predicted probabilities was equal to 0.5. The team defined ungradability for predicting glaucoma as the logarithm of the p-value for the Wilcoxon test.

4.4.14 Mirazzak [195]

Team *Mirazzak* used an ensemble of ConvNeXts [196] and a vision transformer for the screening performance task.

For the ungradability task, they employed the *regret* function, which was proposed by Bibas *et al.* [197] as the generalization error of an explicit expression of the predictive normalized maximum likelihood learner. If the value of *regret* function was high, the samples were considered OOD and they were marked as ungradable.

4.5 Results

This section presents the glaucoma screening performance and robustness of the fourteen participating teams. The final rankings and mean positions of the teams are shown in the first plot of Fig. 4.2. Four teams shared a rank with another team, since their mean positions were exactly equal, causing there to be two teams for each of the ranks #2 and #11.

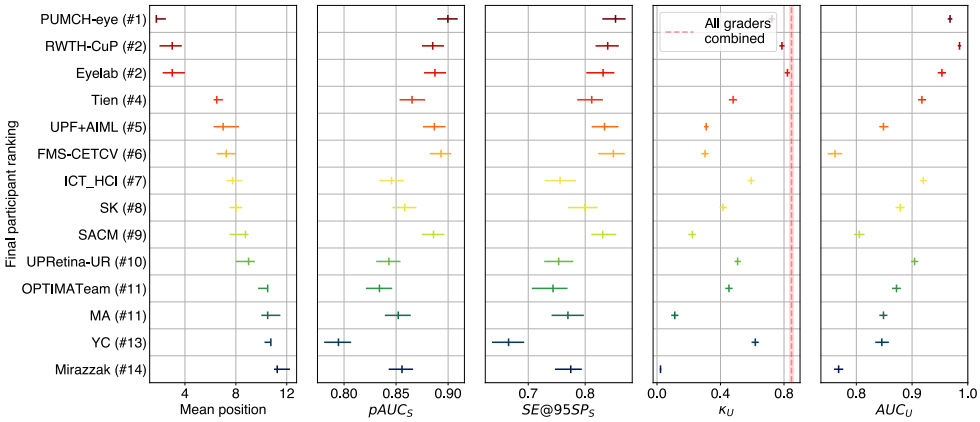


Figure 4.2: Final rankings of all participating teams. The teams are sorted by their final ranking and therefore also by their mean position. The mean position is shown in the left plot and the four challenge metrics are shown in the other four plots. The κ_U of all human graders is indicated with a red dotted line. The width of the horizontal lines in all plots and the shaded area in the plot for κ_U are 95% CIs. We consistently use the same colors to refer to teams in other figures in this manuscript.

4.5.1 Glaucoma screening performance

The glaucoma screening performance of the participating teams is summarized in Fig. 4.2, showing $pAUC_S$ and $SE@95SP_S$ in the second and third plot, respectively. The highest scores for $pAUC_S$ and $SE@95SP_S$ were 0.90 (95% CI: 0.89 – 0.91) and 0.85 (95% CI: 0.83 – 0.87), respectively. These scores were both achieved by team *PUMCH-eye*.

Fig. 4.3a and Fig. 4.3b show the $pAUC_S$ and $SE@95SP_S$ for the ensembles when averaging the RG likelihood output O_1 of the best M participants in terms of the relevant metric. An optimal $pAUC_S$ of 0.91 (95% CI: 0.90 – 0.92) was achieved at $M = 3$. At $M = 2$, an optimal value for $SE@95SP_S$ was reached, which was 0.87 (95% CI: 0.85 – 0.89).

Fig. 4.4a shows the partial ROC curves between 90% and 100% specificity for all participants. The plot also presents the sensitivity and specificity of the human graders with a 95% CI. These were 0.86 (95% CI: 0.84 – 0.87) and 0.94 (95% CI: 0.94 – 0.95), respectively.

In Fig. 4.5, we compare the performance on the REFUGE test set of the final AIROGS algorithms, which were trained on the AIROGS train set, to the performance of the algorithms that were submitted to the REFUGE challenge, which were trained on the REFUGE train set. The top three participants of the REFUGE algorithms achieved AUCs of 0.99, 0.98 and 0.96. For the AIROGS algorithms, the best three AUCs were 0.98, 0.97 and 0.97. The mean $\pm$ std. dev. AUC of all REFUGE and AIROGS algorithms were 0.94 $\pm$ 0.04 and 0.95 $\pm$ 0.02, respectively. Fig. 4.6 presents the relation between the two glaucoma screening performance metrics of all participating AIROGS algorithms on the AIROGS test set and that

performance on the REFUGE test set. For both metrics, almost all AIROGS algorithms (except for team *PUMCH-eye* for $SE@95SP_S$) scored higher on REFUGE than on AIROGS. Of all AIROGS participants, the best $pAUC_S$ and $SE@95SP_S$ on REFUGE were 0.94 and 0.88, respectively.

In Fig. 4.7, the relation between the glaucoma performance of all participating AIROGS algorithms on the AIROGS test set and that performance on GAMMA is shown. For both screening metrics, all AIROGS algorithms scored higher on GAMMA than on AIROGS. Of all AIROGS participants, the best $pAUC_S$ and $SE@95SP_S$ on GAMMA were 1.0 and 1.0, respectively.

4.5.2 Robustness

The robustness metrics of the participating teams are summarized in Fig. 4.2, showing κ_U and AUC_U in the fourth and fifth plot, respectively. The highest scores for κ_U and AUC_U were 0.82 (95% CI: 0.80 – 0.84) and 0.99 (95% CI: 0.98 – 0.99), respectively. These scores were achieved by team *Temirgali* and *RWTH-CuP*, respectively.

Fig. 4.3c and Fig. 4.3d show the κ_U and AUC_U for the ensembles when averaging output O_3 and output O_4, respectively, of the M best algorithms in terms of these respective metrics. An optimal κ_U of 0.85 (95% CI: 0.84 – 0.86) was achieved at $M = 6$. Also at $M = 6$, an optimal value for AUC_U was reached, which was 0.99 (95% CI: 0.99 – 0.99).

In Fig. 4.4b, ROC curves for robustness are shown for all participants. The plot also presents the sensitivity and specificity for separating ungradable from gradable images of the human graders with a 95% CI. These were 0.95 (95% CI: 0.94 – 0.96) and 0.97 (95% CI: 0.97 – 0.97), respectively.

The results on the external DRIMDB dataset are shown in Fig. 4.8, indicating the relation between the ungradability metrics κ_U and AUC_U of all participating AIROGS algorithms on DRIMDB and those metrics on AIROGS. Of all AIROGS participants, the best κ_U and AUC_U on DRIMDB were 0.94 and 1.0, respectively.

4.5.3 Inference time

The time that each algorithm took to perform inference on the test set of the *Final Test Phase* is shown in Fig. 4.9. Please note that these results reflect the total time it took to run the Docker containers provided by the participants. The time it took for the software to load the model weights and to run other setup code defined by the teams was also included in this analysis. Since the test set was split into 38 separate chunks, this initialization and setup code was run at least 38 times for each participating team, as well.

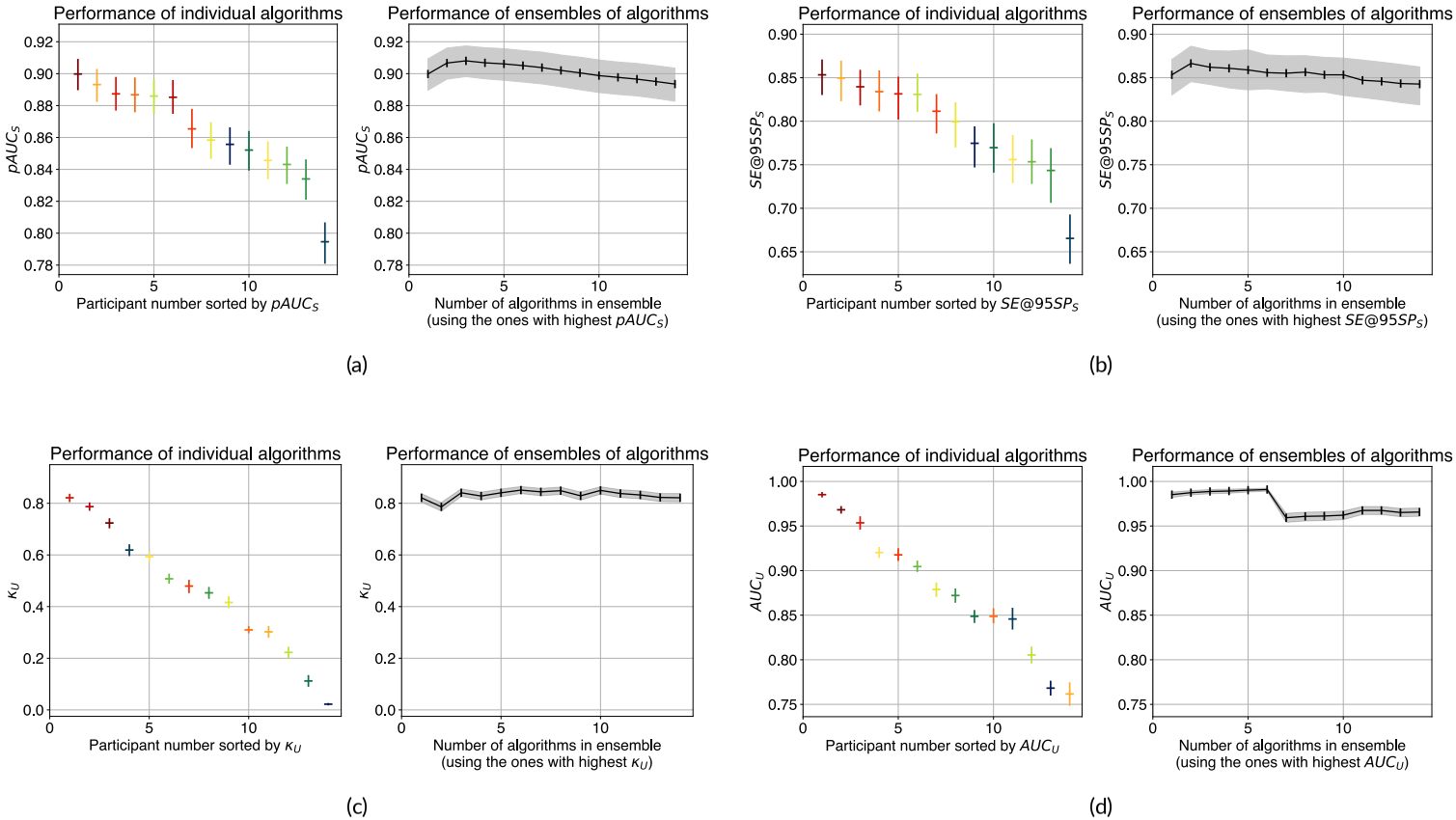


Figure 4.3: The four challenge metrics **(a)** $pAUC_S$, **(b)** $SE@95SP_S$, **(c)** κ_U, and **(d)** AUC_U for the ensembles generated by incrementally fusing one algorithm at a time. The algorithms were fused by averaging the outputs of all algorithms in the ensemble. The vertical lines in the left plot and the shaded areas in the right plots indicate 95% CIs.

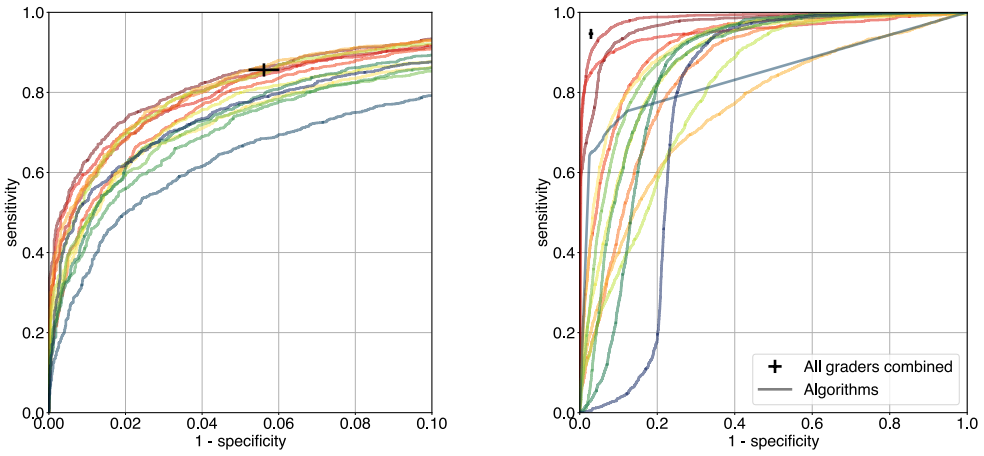


Figure 4.4: ROC curves for both challenge tasks. The sensitivity and specificity of all human graders on the AIROGS test set combined are indicated with black lines. Respectively, the width and height of the black horizontal and vertical lines are 95% CIs. In **(a)**, the partial ROC curve (90%-100% specificity) for screening is shown, with 1,602 positive (RG) and 8,134 negative (NRG) images from the AIROGS test set. In **(b)**, the ROC curve for robustness is shown with 1,554 positive (ungradable) and 9,736 negative (gradable) images from the AIROGS test set.

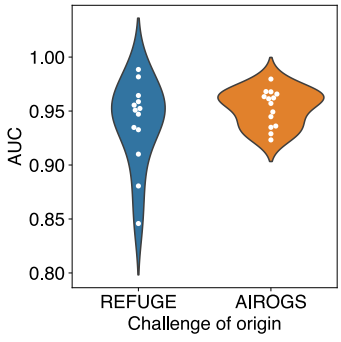
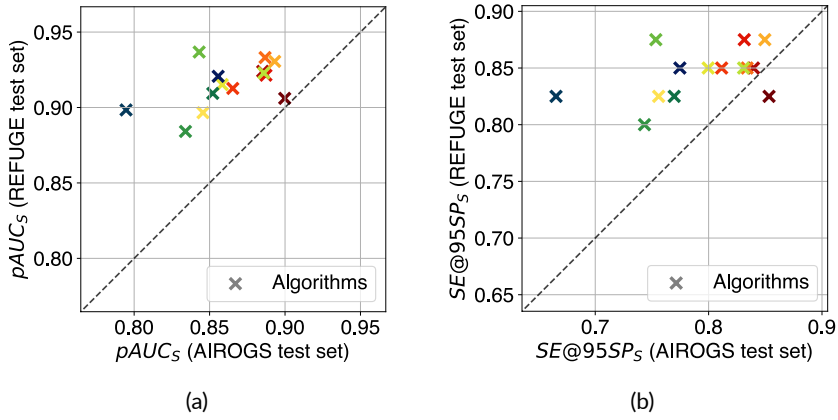


Figure 4.5: Comparison of the AIROGS and REFUGE algorithms, tested on the REFUGE test set, visualized as violin and swarm plots. The final algorithms that were developed for the REFUGE challenge itself and for the AIROGS challenge are shown on the left and right, respectively. The AIROGS algorithms were only trained on the AIROGS train set and were not retrained with the REFUGE dataset.

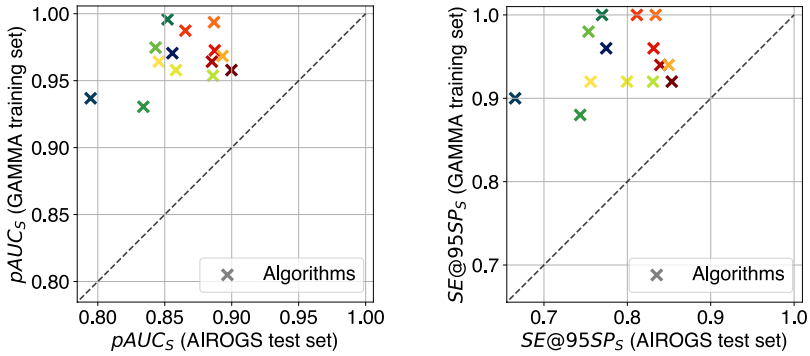
4.6 Discussion

AI models have been shown to be effective at detecting glaucoma in CFPs, but most studies lack evidence of robustness to real-world scenarios in which unexpected OOD data can be presented due to various causes. To this end, we relied on the community to develop robust AI solutions for glaucoma screening based on the largest multi-center real-world



(a) (b)

Figure 4.6: Performance of the participating AIROGS algorithms on the REFUGE dataset, compared to their performance on the AIROGS dataset. Both screening metrics **(a)** $pAUC_S$ and **(b)** $SE@95SP_S$ are shown.



(a) (b)

Figure 4.7: Performance of the participating AIROGS algorithms on the GAMMA dataset, compared to their performance on the AIROGS dataset. Both screening metrics **(a)** $pAUC_S$ and **(b)** $SE@95SP_S$ are shown.

CFP dataset with glaucoma labels. We organized the AIROGS challenge around this dataset, ensuring the resulting algorithms are reusable in a cloud-based environment. We applied these algorithms to ungradable data, while the participants could only train on gradable data to ensure robustness to any kind of ungradable data, and to other publicly available datasets to assess their generalization.

4.6.1 Overall findings

The team with the highest $SE@95SP_S$ scored expert-level screening performance on the AIROGS test set with a sensitivity of 0.85 (95% CI: 0.83 – 0.87) at 95% specificity, similar to the sensitivity of 0.86 (95% CI: 0.84 – 0.87) at a specificity of 0.94 (95% CI: 0.94 – 0.95) of

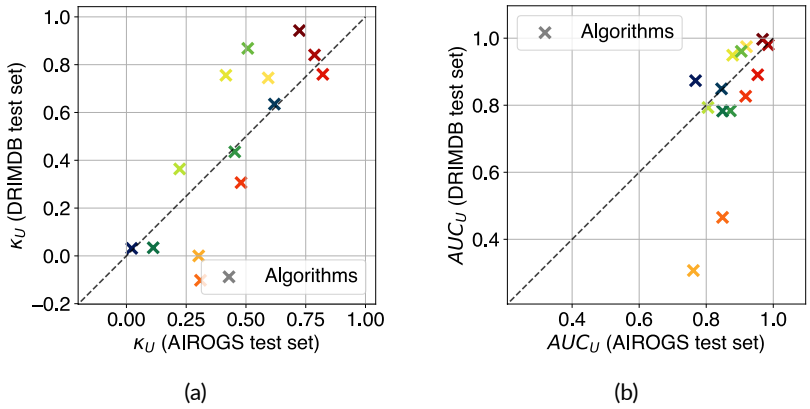


Figure 4.8: Performance of the participating AIROGS algorithms on the DRIMDB dataset, compared to their performance on the AIROGS dataset. Both robustness metrics **(a)** κ_U and **(b)** AUC_U are shown.

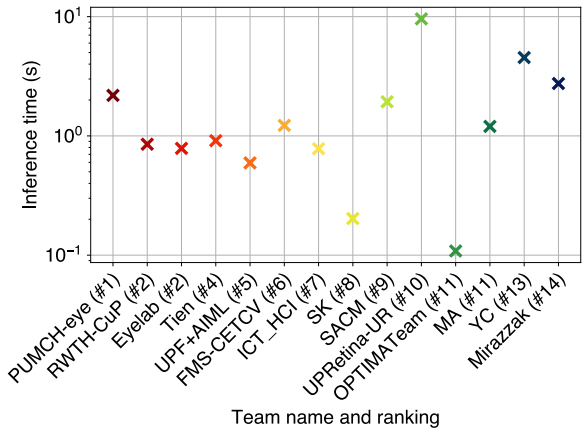


Figure 4.9: Average inference time per CFP in the test set of the *Final Test Phase*. This time includes the actual inference time, model initialization, and other setup code executed by the submitted Docker containers.

human graders. The highest $pAUC_S$ that was achieved by any of the teams was 0.90 (95% CI: 0.89 – 0.91). Ensembling the different participating methods improved the screening performance even further, to 0.91 (95% CI: 0.90 – 0.92) and 0.87 (95% CI: 0.85 – 0.89) for $pAUC_S$ and $SE@95SP_S$, respectively. Our analysis revealed that ensembling improved the performance for all metrics up to a certain point at which adding further models to the ensemble resulted in a decline in performance. A probable reason for this is that the models added after this point under-perform to such a degree that their outputs negatively impact the performance of the ensemble, instead of improving it. Seven out of fourteen teams exceeded the minimum performance of 80% sensitivity and 95% specificity that was required by human graders who were periodically monitored during the grading pro-

cess. This shows these models can provide similar performance to human graders for glaucoma screening, suggesting that AI can potentially play a role in an automated screening process.

We also evaluated the screening performance of the algorithms on two external test sets. Even though the algorithms were trained on AIROGS data, they achieved very high performances on the two external test sets, showing reproducible results in different sets and populations. On average, the participating AIROGS algorithms scored slightly higher on the REFUGE dataset than the REFUGE participants. We found that the participating algorithms scored substantially higher on these external datasets than on the AIROGS test set, indicating the value of a challenging real-world dataset. This strong generalization of the developed solutions also shows the potential of models trained on our dataset to be successfully implemented in screening programs with limited to no loss of performance. Unlike the external datasets, the AIROGS dataset represents a screening population, which likely consists of a relatively large amount of individuals with lower severity levels of glaucoma compared to clinical populations. Since more severe cases are expected to be picked up easier than less severe cases, the underlying ratio between less and more severe positive glaucoma cases could be a cause of the observation that the external test set performance was higher than the internal test set performance.

The robustness to ungradable data in the AIROGS test set was evaluated for each team using the metrics κ_U and AUC_U. The teams that performed the best in terms of these metrics achieved 0.82 (95% CI: 0.80 – 0.84) and 0.99 (95% CI: 0.98 – 0.99) for κ_U and AUC_U, respectively. Human experts did reach a higher κ_U of 0.85 (95% CI: 0.84-0.86) for this task. Moreover, they achieved a sensitivity of 0.95 (95% CI: 0.94 – 0.96) and a specificity of 0.97 (95% CI: 0.97 – 0.97) for detecting ungradable cases, while the team with the best AUC_U achieved a lower sensitivity at 97% specificity of 0.90 (95% CI: 0.88-0.92). Although the teams achieved relatively high performances, they still achieved lower performance at the robustness task than human experts. This shows this task was especially challenging, possibly because the participating teams could not use ungradable development data or because their robustness approaches focused on specific forms of ungradability.

We also assessed robustness on the external DRIMDB dataset. The best-scoring team on this dataset scored very high performances; they achieved 0.94 and 1.0 for κ_U and AUC_U, respectively. These two metrics were lower on the AIROGS dataset for that team. This also indicates very strong generalization to other datasets for the robustness task. The high ungradability detection performance also indicates robustness to other diseases in the image, as diabetic retinopathy was prevalent in the gradable subset of DRIMDB and the best algorithms did not classify these diseases as ungradable.

A large difference in performance between participating teams can be observed, both for the screening and the robustness task. Therefore, we think it is important to identify which methodological choices were made predominantly by top-performing teams. One of the most notable differences between the top three participants and the rest was the use of transformers. Outside of the top three, only the latest-placed team used a transformer. One of the possible reasons for this superiority that is achieved by transformers compared to CNNs could be their effectiveness at modeling long-range dependencies [198,199]. This allows for a better understanding of contextual information, which is generally believed to be beneficial in medical imaging [199]. Empirically, transformers and methods that combine transformers with CNNs have previously also been shown to outperform CNNs in medical image analysis [200–202], which is in line with our findings.

Moreover, all best three participants manually labeled ODs for training either a segmentation or detection model to crop around the OD during pre-processing. Even though this was also done by two other teams, this seems like an effective strategy to achieve higher screening performance. A likely reason for the effectiveness of this approach is that most glaucoma-related imaging features can be found on or around the OD. This shows how a priori medical knowledge could still be of value even when a large amount of data is available. A less important factor appears to be the number of manually labeled ODs. A possible reason for this could be that the OD detection or segmentation network is not required to be extremely accurate, as combining a rough localization of the OD with a large enough padding margin could also suffice to crop the image during pre-processing.

Since the development set only consisted of images that were labeled gradable (either RG or NRG) and the use of external fundus data was prohibited, all teams came up with an uncertainty or OOD detection method based on the gradable data for the robustness task. The ungradability methods of the top three participants in terms of mean position, κ_U, and AUC_U, all revolved around the confidence of a neural network that localized the OD. Of the other participants, only the ninth-placed team had such an approach. Apart from these methods based on OD detection, only team *UPF+AIML* implemented a different robustness technique that was also based on domain knowledge. This raises the impression that solutions based on domain knowledge are more effective for robustness than more general OOD detection solutions. However, it still needs to be evaluated if such approaches are robust for other general tasks (not glaucoma screening) or other sources of OOD data.

For calculating the κ_U metric, the participants were required to output a binary decision on ungradability. A popular approach, especially among the top participants, was to manually identify relatively low-quality images in the development set and base a threshold for this binary output on that subset. This technique was employed by the best three,

fifth, tenth, and twelfth teams in terms of κ_U. This indicates that this could be a successful approach, although not in general as the accuracy of this binary value is also highly dependent on the quality of the scalar output for ungradability O_4 that is being thresholded. We found the difference between the ranking in terms of κ_U and AUC_U of one team, in particular, stood out. Team *YC* ranked only eleventh for AUC_U (which depended on the scalar output O_4), but ranked fourth in terms of κ_U (which depended on the binary output O_3), indicating the approach they used for thresholding their scalar value was highly effective. The difference between their AUC_U and κ_U was 7, while the next biggest value of this difference was only 3. Team *YC* indeed came up with a relatively sophisticated method for binarizing O_4 compared to others, based on a Wilcoxon one-sample test to statistically test whether the mean of the predicted probabilities from a Monte-Carlo drop-out approach was 0.5 or not.

4.6.2 Strengths and limitations

The dataset presented in this paper substantially exceeds what was publicly available before in terms of number of images and patients. The dataset is also highly diverse because of the large number of different sites, cameras that were used, and ethnicities. Comorbidities were not excluded from the dataset, as our goal was to develop tools that are robust to data with such conditions in real-world screening settings. These conditions should not have an influence on the classification of glaucoma or ungradability. The quality of the labels was controlled by the initial and periodical evaluation of human graders, the fact that each image was independently labeled twice by two trained graders, and, in case of disagreement, by a highly experienced reader. The participants submitted their solutions as containerized algorithms, allowing reproducibility, facilitating inference on other data, and preventing manual manipulation of the test set.

One of the rules of the AIROGS challenge was the prohibition of the use of external fundus data for development. A limitation of this work is the fact that we cannot be sure if any of the teams used such data in their development process. A possible approach to prevent this and make the process fairer is to have participants submit a containerized algorithm for training, which would be trained by the challenge organizers with private challenge training data. Nevertheless, with such an approach it would still be challenging and time-consuming for the challenge organizers to verify if the training containers do not contain any weights pre-trained on other data.

The teams that participated in the competition were permitted to create their own manual annotations on the data and employ them in the development of their models. We made a deliberate decision not to forbid this practice, as we believed that the possibility of

achieving superior results outweighed the disadvantage of potentially introducing a slight unfairness due to some teams having access to larger manual labor workforces than others. Teams *PUMCH-eye*, *RWTH-CuP*, *Eyelab*, *UPF+AIML*, *ICT_HCI*, *SACM*, and *YC* took advantage of this opportunity and produced at least one of the subsequent manual annotations: OD detection or segmentation, vessel segmentation, and identification of low-quality images in the development set. As a result, the final rankings of the teams have likely been influenced by this practice, and it is important to consider this potential bias when comparing solutions. We think it is also worth noting that the aforementioned solution for promoting fairness of submitting a containerized algorithm for training on private data would disable manual annotation of development data.

The dataset used for the challenge is diverse, but improvements could still be made in that respect. All screening sites were based across the United States of America, raising the question of whether a more generalizable model could be obtained with data from across the world. On the other hand, we showed that many algorithms trained on the AIROGS dataset performed at least as well on three external test sets, of which two originated from China and one from Turkey, as on our internal test set.

Not all research groups working in the field of retinal image analysis participated in this challenge and many teams that joined the challenge did not submit a solution to the *Final Test Phase*. Possible reasons for this include that many teams saw their results did not match to ones already present on the leaderboard, that the barrier for some teams was too high to get a solution wrapped in a Docker container, or that they were not able to finish in time. Therefore, we would like to stress the challenge is still open and we are curious to see if the community can make further improvements. After all, especially for the robustness task, there seems to be room for improvement, given the gap with the human grader performance.

4.6.3 Future directions

Based on the solutions that were presented by the teams, we think it would be valuable to combine methodologies from different participants and to work further on their ideas. For example, as we mentioned before, team *YC* apparently had a highly effective method for thresholding their ungradability scores as their κ_U was very high compared to their AUC_U. A possible future direction would be to combine methods of high performance in terms of AUC_U with the binarization technique from team *YC*. Moreover, we observed that algorithms that scored high in terms of robustness, used domain knowledge for this aspect of the challenge. Possible future directions could be to explore other ways to incorporate domain knowledge into an ungradability method. This observation also leads to

the question of whether there are more fields in medical image analysis in which domain knowledge can be leveraged for uncertainty estimation and OOD detection.

Next to a decision on RG and NRG presence, the graders were asked to provide which clinical, glaucomatous features were present in the eyes they classified as RG, as listed in Section 4.2.1 and further described by [74]. This information was not yet included in the dataset release for this challenge, as it fell outside the scope of this challenge. Future solutions and challenges could be developed with this information, possibly resulting in more explainable algorithms.

This challenge only focused on classification based on a single CFP. It may be interesting to explore the effect on screening performance and robustness of including various types of metadata in our dataset, which we have available but have not been published yet. This metadata, although missing for some images, includes the camera type, age, and anonymous patient identification (which can be used to link two eyes to a single patient).

In order to ensure the safe and effective implementation of the AI models for glaucoma screening described in this paper, several important steps need to be undertaken. González-Gonzalo *et al.* [203] provide valuable insights into the key aspects that are crucial for the integration of AI models in ophthalmic practice.

Among these aspects, additional retrospective validation studies play a significant role in validating the performance and generalization of these models. An external evaluation with substantially different data was already performed in this study. However, we think it is crucial to evaluate with additional large screening datasets that represent real-world scenarios before practical implementation. Prospective validation studies and cost-effectiveness analyses are also essential for evaluating the accuracy, reliability, and generalization of glaucoma screening AI models in real-world settings. These analyses are especially important for screening programs since screening solutions that are not specific enough can have substantial negative financial impacts, as they can lead to unnecessary hospital visits. It is also essential to identify and mitigate potential limitations such as data quality, which the AIROGS challenge aimed to address, model interpretability, integration with screening workflows, and potential biases. Finally, establishing mechanisms for post-market surveillance is vital to monitor and evaluate the performance and safety of these AI models after their regulatory approval.

Further implementation and real-world evaluation of these algorithms are needed, as described above. As this was considered out of the scope of the current manuscript, we leave the execution of these steps to future work.

4.7 Conclusions

We presented the results of community-acquired algorithms tested on real-world data for robust glaucoma screening from CFP. The best algorithms performed similarly in terms of screening to the carefully trained and selected human graders, and were shown to be effective at flagging images that could not be graded. Methodological choices predominantly made by the best teams included, for the screening task, the use of vision transformers and the incorporation of optic disc detection models in pre-processing and, for the robustness task, out-of-distribution detection approaches based on domain knowledge. We hope the unprecedented size and real-world nature of the dataset we released and the algorithms that were developed using this dataset will help towards implementing robust AI for glaucoma screening.

Acknowledgements

This research was funded in part by Eurostars grant E12712 and supported in part by Amazon Web Services.

PART III

Mitigating performance drops under data shift

5

Reliable classification using Uncertainty-Based Instance eXclusion

Authors: Coen de Vente, Bram van Ginneken, Carel B. Hoyng, Caroline C. W. Klaver, Clara I. Sánchez

Original title: Uncertainty-Aware Multiple-Instance Learning for Reliable Classification: Application to Optical Coherence Tomography

Published in: Medical Image Analysis (Volume: 97, October 2024)

https://doi.org/10.1016/j.media.2024.103259

Abstract

Deep learning classification models for medical image analysis often perform well on data from scanners that were used to acquire the training data. However, when these models are applied to data from different vendors, their performance tends to drop substantially. Artifacts that only occur within scans from specific scanners are major causes of this poor generalizability. We aimed to enhance the reliability of deep learning classification models using a novel method called Uncertainty-Based Instance eXclusion (UBIX). UBIX is an *inference-time* module that can be employed in multiple-instance learning (MIL) settings. MIL is a paradigm in which instances (generally crops or slices) of a bag (generally an image) contribute towards a bag-level output. Instead of assuming equal contribution of all instances to the bag-level output, UBIX detects instances corrupted due to local artifacts on-the-fly using uncertainty estimation, reducing or fully ignoring their contributions before MIL pooling. In our experiments, instances are 2D slices and bags are volumetric images, but alternative definitions are also possible. Although UBIX is generally applicable to diverse classification tasks, we focused on the staging of age-related macular degeneration in optical coherence tomography. Our models were trained on data from a single scanner and tested on external datasets from different vendors, which included vendor-specific artifacts. UBIX showed reliable behavior, with a slight decrease in performance (a decrease of the quadratic weighted kappa (κ_w) from 0.861 to 0.708), when applied to images from different vendors containing artifacts; while a state-of-the-art 3D neural network without UBIX suffered from a significant detriment of performance (κ_w from 0.852 to 0.084) on the same test set. We showed that instances with unseen artifacts can be identified with OOD detection. UBIX can reduce their contribution to the bag-level predictions, improving reliability without retraining on new data. This potentially increases the applicability of artificial intelligence models to data from other scanners than the ones for which they were developed. The source code for UBIX, including trained model weights, is publicly available through `https://github.com/qurAI-amsterdam/ubix-for-reliable-classification`.

5.1 Introduction

Deep learning models for medical image analysis applications are often trained on data that is acquired with one or a selected number of scanner types and/or acquisition protocols. When applying these trained models on data from different scanners or protocols, the performance tends to plummet [204, 205]. This negatively affects the reliability of these systems, which is a main aspect of trustworthy AI [206, 207], and its wide integration and adoption in clinical practice. In general, convolutional neural networks (CNNs) are known to fail when they are applied under dataset shift or to out-of-distribution (OOD) datasets; and approaches to address this effect are being investigated [208]. This OOD nature of the data occasionally only stems from local areas in images, such as local artifacts. These local artifacts occur frequently in data from specific vendors or particular scanning protocols, and can be found in multiple medical imaging fields, such as contrast-enhanced mammography data [209] and optical coherence tomography [210]. In images with these types of artifacts, there generally are sufficient parts in a sample which are in-distribution (ID) to form a correct prediction if the model would in some way only focus on those parts of the data and neglect the OOD areas.

To achieve this increased robustness to local OOD areas in images, we propose Uncertainty-based Instance eXclusion (UBIX). This approach builds upon multiple-instance learning (MIL), a form of weakly supervised learning popular in medical image analysis [211,212]. In MIL, a labeled bag (usually the whole input image) consists of multiple unlabeled instances (usually image patches, regions or slices). During deep MIL, instances are considered individually by a neural network, and the instance-level outputs, each contributing equally, are combined to obtain a bag-level prediction using a MIL pooling function [213]. Instead of assuming equal contribution, the UBIX approach assumes that some of the instances might be corrupted due to local artifacts, identifies these instances on-the-fly using uncertainty estimation, and reduces or ignores its contribution to the bag-level prediction using the so-called UBIX function before the MIL pooling function (see Fig. 5.1). To the best of our knowledge, this is the first method that uses OOD detection in such a manner to increase reliability.

Although our method is applicable to any instance definition (such as 2D or 3D patches) or bag definition, we focus on MIL problems in which slices are instances and full 3D volumes are bags. Specifically, we focus on classifying age-related macular degeneration (AMD) in optical coherence tomography (OCT), as there is a plurality of manufacturers and scanner versions in the field of OCT. [56] list fourteen companies that produce OCT scanners for ophthalmic applications and the type of imaging artifacts that occur can differ sub-

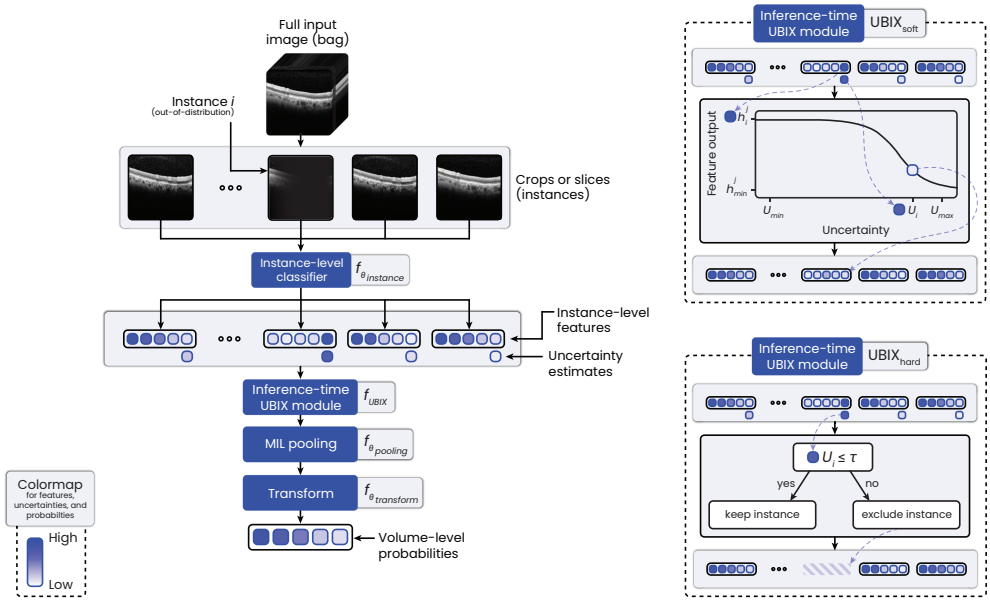


Figure 5.1: Overview of the general MIL pipeline (left) and how the two UBIX variants are integrated into this pipeline (right). Each MIL instance is fed into the same classifier. During inference time, a UBIX function converts pre-UBIX instance-level features based on their respective uncertainties to post-UBIX instance-level features. The instance-level post-UBIX features are then converted to bag-level outputs using MIL pooling. During training, the pre-UBIX features are fed directly into the MIL pooling function. The thin dashed blue arrows indicate the flow of one example feature and its associated uncertainty. h_i^j is the instance-level feature value for instance i and feature j. h_{min}^j is the minimum instance-level feature value in the validation set for feature j. U_i is the uncertainty associated with instance i. U_{min} and U_{max} are the minimum and maximum uncertainty in the validation set, respectively.

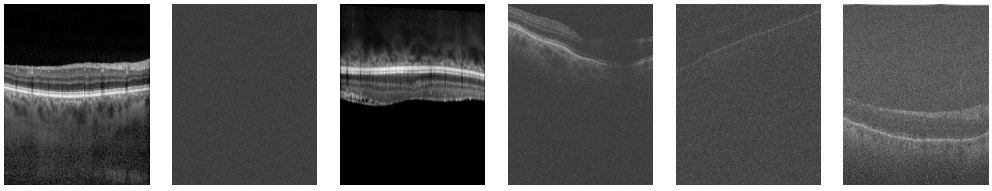


Figure 5.2: Examples of real imaging artifacts in OCT. Subfigure **(a)** shows a normal B-scan without imaging artifacts for reference. The other subfigures show various common imaging artifacts. The images **(a)** and **(c)** originate from the H_{val} dataset, while images **(b)**, **(d)**, and **(e)** originate from the T_{test} dataset, and image **(f)** originates from the B_{test} dataset. **(a)** No imaging artifacts. **(b)** Blinking artifact. **(c)** Vertically flipped. **(d)** Shadow. **(e)** Noise. **(f)** Noise.

stantially across scanners [210]. These artifacts include slices (B-scans) that are fully black due to blinking, vertically flipped B-scans, shadows, and noise. For example, blinking artifacts and very noisy B-scans are much less common in certain scanners, specifically ones

with higher speed and eye tracking software [210], due to internal B-scan averaging and rescanning strategies [214]. A number of examples of imaging artifacts occurring in OCT are shown in Fig. 5.2.

While our primary investigation centers on UBIX's assessment in ophthalmic applications, its potential utility spans diverse medical image tasks. UBIX aims to increase the robustness to local imaging artifacts, which may be under-represented during training but at least sporadically manifest in clinical settings. A relevant example pertains to the observation from [215] of small colon tissue regions in a histopathology prostate training set, leading to locally OOD instances. Furthermore, various artifacts in histopathology, such as foreign objects, scratches, elastic deformations, and fingerprints, have been outlined by [216]. These artifacts can result from natural origins, preprocessing, and digitization steps. [217] discussed metal artifacts in CT scans. Commercially available CT scanners employ a range of techniques to address metal artifacts, leading to varied appearances of these artifacts in scans [217]. In these CT and histopathology examples, instances could be defined as crops (volumetric crops in CT, and 2D crops in histopathology), while the whole slide image and CT volume could be the bags. We expect that classification models developed in these contexts may potentially benefit from UBIX as well.

We evaluate the generalizability of our proposed models by training on data acquired with a scanner from one vendor, while evaluating with data from scanners of other vendors. We show that UBIX increases this generalizability using a baseline comparison. Moreover, we systematically analyze the ability of UBIX to detect OOD instances by gradually introducing artificial image artifacts that occur naturally as well. The trained algorithm is publicly available for inference on the online platform of Grand Challenge*.

5.2 Related work

5.2.1 Multiple-instance learning in medical imaging

One of the most common medical application in which MIL is applied is histopathology [218–221], mainly because it is very labor-intensive and time-consuming to manually annotate entire whole slide images on instance-level. [212] used an attention-based MIL pooling layer and evaluated it on an MNIST-based dataset and histopathology datasets. Other medical modalities to which MIL with deep learning has been applied include ultrasound [222, 223], computed tomography [224, 225] and magnetic resonance imaging [226, 227].

Several methods have been proposed to integrate confidence or uncertainty estima-

*https://grand-challenge.org/algorithms/amd-classification-in-oct-with-ubix/

tion at instance-level in MIL approaches. For example, Integrated Instance-Level and Bag-Level MIL (IIB-MIL) [228] is a MIL approach that integrates instance-level and bag-level supervision, using a frozen instance-level encoder pre-trained with weak supervision. After pre-training, it optimizes the instance-level features with a label-disambiguation module, which incorporates a confidence bank, aiming to mitigate the effect of training with noisy labels. This confidence bank is a form of uncertainty estimation but serves a different purpose in their pipeline than the uncertainty estimation in UBIX does.

Weakly supervised knowledge distillation (WENO) [229] is a student–teacher framework that consists of an instance-level classifier (the student) and a bag-level classifier (the teacher). Attention-based scores from the teacher are used to train the student. It also uses hard positive instance mining, which relies on a form of confidence estimation to find hard examples. In WENO, this ensures that the model not only learns to define a positive bag from easy positive instances but also predicts harder positive instances as positive. Even though [229] show this is an effective strategy to increase model performance, they do not aim to increase robustness to OOD instances. This is in contrast with the aim of UBIX.

Unlike our proposed method, these approaches may suffer from limited robustness when transferred to other distributions, as they do not explicitly employ methods that aim to improve robustness when transferring to OOD data.

5.2.2 Out-of-distribution detection

OOD detection is the identification of samples that originate from a different distribution than the training distribution. Such samples generally have high model predictive uncertainties, given a good uncertainty estimation method. [28] proposed a simple baseline for OOD detection using the maximum class probability as confidence scores. Another early work was Monte Carlo dropout (MC-DO), in which they leveraged dropout to estimate uncertainty [194]. [208] compared a number of methods for OOD detection and uncertainty estimation including MC-DO. They found that deep ensembling [230] was one of the top-performing methods for OOD detection. Since then, other popular methods for uncertainty estimation and OOD detection have been published [181, 231, 232].

Uncertainty estimation has been investigated for medical image analysis as well, such as [233], who used deep ensembles to calibrate probabilities in segmentation maps. [234] detected incorrect orientation or anatomy in X-rays using an OOD detection metric called FRODO, defined as the Mahalanobis distance of test samples to samples in the train set. Furthermore, [215, 235] used multi-head CNNs, an approach similar to deep ensembles, to detect images with lymphoma in histopathology as OOD samples.

In general, uncertainty estimation in medical images are used as an additional output to assess the behavior of the developed models or to identify abnormalities as OOD samples. In contrast, our proposed approach takes into account OOD detection during inference to increase classification robustness against data shift.

5.2.3 Robustness against data shift in OCT

Related works have successfully applied machine learning methods for AMD classification from OCT, but did not specifically focus on robustness to OOD data [236–241].

The following works studied robustness against data shift in OCT. [205] used an OCT segmentation network of which the output was fed into a classification network, which in turn outputted a referral suggestion, diagnosis probabilities for multiple retinal disease features, such as choroidal neovascularization (CNV) and geographic atrophy (GA), and volume estimations of drusen and epiretinal membranes. The error rate on their internal test set with OCTs from the same scanner, as the development set, *i.e.*, Topcon, was 5.5%, but the error rate increased to 46.6% when transferred to an external set with OCTs from a different scanner, *i.e.*, Heidelberg Spectralis. When retraining their segmentation network with data from this scanner, the error rate improved to 3.4%. [242] and [243] used a CycleGAN to transform OCT scans acquired on a device that was not used during training to have a similar appearance as the training data. For retinal fluid [242, 243] and layer [243] segmentation, they observed a generalizability improvement when applying this domain adaptation technique, compared to traditional transformation strategies.

The main downside of these methods is their requirement for – albeit annotated or not – data from the new setting. Acquiring and annotating this new data, as well as any potential retraining, is a time-consuming and expensive process. Moreover, if these models are unknowingly applied in settings that are highly different from the development setting, models can fail silently, potentially causing misdiagnoses. We propose a method that reduces the performance drop when a model is transferred to a setting unlike its development setting, without the requirement for acquiring or labeling data originating from this new setting.

5.3 Methods

Our paper introduces UBIX, a method that enhances the robustness of classification models to locally OOD data. Specifically, our approach is designed for scenarios where a portion of a given sample, such as a crop or slice of an image, differs from the training data distribution. UBIX is implemented in the MIL paradigm and is *solely applied during inference*. It

can theoretically be applied seamlessly to any existing MIL model that has already been trained. In Section 5.3.1, we provide an introduction to MIL and related concepts. UBIX relies on the identification of local OOD areas in the input sample, by performing uncertainty estimation on MIL instances. In Section 5.3.2, we present various state-of-the-art uncertainty estimation methods and propose variants of uncertainty measures that are optimized for ordinal labels. Finally, Section 5.3.3 elucidates how UBIX works and how it integrates instance-level uncertainty estimation into MIL models.

5.3.1 Multiple instance learning (MIL)

The MIL paradigm

MIL is a form of supervised learning designed to address scenarios where our dataset is organized into bags of instances. Each bag, denoted as $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_I\}$, can contain a varying number of instances, and each bag is associated with a bag-level label, $Y \in \{1, \ldots, C\}$, where C represents the number of possible classes. Individual instances within a bag do not have explicit labels. During training, only bag-level labels are available.

MIL employs an instance-level classifier, which transforms each instance $\mathbf{x}_i$ into a feature vector $\mathbf{h}_i \in \mathbb{R}^J$ using a function parameterized by $\theta_{instance}$, represented as $\mathbf{h}_i = f_{\theta_{instance}}(\mathbf{x}_i)$. There are two primary types of MIL approaches: instance-based and embedding-based. In the instance-based approach, J is equal to C, which means that the instance-level classifier produces instance-level logits for each class. In contrast, the embedding-based approach allows J to represent the number of features, potentially using a higher-dimensional space.

Once the instance-level classifier has been applied, MIL pooling is employed to aggregate these vectors into a bag-level feature representation, denoted as $\mathbf{h}_X$, using a function parameterized by $f_{\theta_{pooling}}$. More formally, $\mathbf{h}_X = f_{\theta_{pooling}}(\mathbf{h}_1, \ldots, \mathbf{h}_I)$. In most MIL pooling functions, $\mathbf{h}_X \in \mathbb{R}^J$. However, when distribution pooling [244] is used, a marginal feature distribution is estimated instead of a single scalar value for each feature. This case is elaborated upon further in Section 5.3.1. Following MIL pooling, a bag-level representation transformation function, parameterized by $\theta_{transform}$, is applied to the bag-level feature vector $\mathbf{h}_X$. This transformation maps $\mathbf{h}_X$ to the predicted bag-level label, denoted as $\hat{Y}$, and is expressed as $\hat{Y} = f_{\theta_{transform}}(\mathbf{h}_X)$.

In deep learning-based MIL, the instance-level classifier $f_{\theta_{instance}}$ is typically implemented as a neural network. In the instance-based approach, the bag-level representation transformation $f_{\theta_{transform}}$ is the identity function, enabling a direct link between each class in the bag-level outputs and the outputs on the instance-level. In the embedding-

based approach, $f_{\theta_{transform}}$ is parameterized using a neural network. Depending on the chosen pooling function, even the MIL pooling function $f_{\theta_{pooling}}$ may be parameterized using a neural network (see Section 5.3.1).

MIL pooling functions

The MIL pooling function $\theta_{pooling}$ aggregates instance-level feature vectors $\mathbf{h}_1, \ldots, \mathbf{h}_I$, where $\mathbf{h}_i = [h_i^1, \ldots, h_i^J]$ and $h_i^j \in \mathbb{R}$, to a bag-level feature vector $\mathbf{h}_X = [h_X^1, \ldots, h_X^J]$, where $h_X^j \in \mathbb{R}$. A requirement for the MIL pooling function is that it can be applied to any number of instances J. The following common MIL pooling functions all follow this requirement and are also described by [245]:

- **Max pooling:** Max pooling selects the maximum value for each feature across all instances in the bag:
$$h_X^j = \max_{i=1}^{I} h_i^j, \quad \text{for } j = 1, 2, \ldots, J.$$

- **Mean pooling:** Mean pooling computes the mean value of each feature across all instances in the bag:
$$h_X^j = \frac{1}{I} \sum_{i=1}^{I} h_i^j, \quad \text{for } j = 1, 2, \ldots, J.$$

- **Attention pooling [212]:** Attention pooling uses a weighted average of instances, where the weights are determined by a neural network. This approach allows for generating bag-level features based on the importance of each instance. The Attention pooling mechanism is mathematically represented as follows:
$$h_X^j = \sum_{i=1}^{I} a_i h_i^j, \quad \text{for } j = 1, 2, \ldots, J,$$
where a_i are the instance-specific attention weights, calculated as:
$$a_i = \frac{\exp(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_i^T))}{\sum_{k=1}^{I} \exp(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_k^T))}.$$
Here, $\mathbf{w} \in \mathbb{R}^{L \times 1}$ and $\mathbf{V} \in \mathbb{R}^{L \times J}$ are learnable parameters.

- **Distribution pooling [245]:** Distribution pooling estimates marginal feature distributions for each feature dimension. This provides a richer representation of the data, as individual features are defined as distributions instead of single scalar values. The formula for distribution pooling involves estimating these marginal distri-

butions. Given instance-level feature vectors $\mathbf{h}_1, \ldots, \mathbf{h}_I$, the goal is to find a bag-level representation $\mathbf{h}_X = [p_X^1, p_X^2, \ldots, p_X^J]$, and p_X^j is the estimated marginal distribution of the feature j. The estimated marginal distribution p_X^j is calculated using kernel density estimation with a Gaussian kernel having a standard deviation σ:

$$p_X^j(v) = \frac{1}{I}\sum_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(v-h_i^j)^2} \quad \text{for } j = 1, 2, \ldots, J.$$

Additionally, in order to use these distributions as vectors in a neural network, the distributions are binned. These binned representations h_X^j, are obtained by sampling the values v from p_X^j at M equally spaced bins within the range of possible values, $v_b = \frac{b}{M-1}$ for $b = 0, 1, \ldots, M-1$. This results in $h_j^X \in \mathbb{R}^M$ for $j = 1, 2, \ldots, J$, and the binning formula can be expressed as:

$$h_X^j = \left[p_X^j(v = v_b) \,\middle|\, v_b = \frac{b}{M-1},\ b = 0, 1, \ldots, M-1 \right] \quad \text{for } j = 1, 2, \ldots, J.$$

Distribution pooling is flexible and can capture rich information about the distribution of features within a bag. It avoids the loss of information associated with point estimate-based pooling methods.

- **Distribution with attention pooling [245]:** Distribution with attention pooling combines Attention pooling with Distribution pooling. Feature distributions are calculated in the same way for each feature as with distribution pooling, except an attention weight a_i is incorporated:

$$p_X^j(v) = \frac{1}{I}\sum_{i=1}^{I} a_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(v-h_i^j)^2} \quad \text{for } j = 1, 2, \ldots, J.$$

The instance-level weight a_i is calculated in the same way as in Attention pooling and, similarly to distribution pooling, binning is performed to transform these distributions into vectors.

- **TransMIL [246]**: TransMIL is a MIL pooling method that is based on the Transformer architecture [10]. The TransMIL pooling function can be expressed as follows:

$$\mathbf{h}_X = \text{TPT}(\mathbf{h}_1, \ldots, \mathbf{h}_I),$$

where the TPT is a module that consists of two Transformer layers and a position encoding layer. Further details about the architecture can be found in the paper from [246].

5.3.2 Uncertainty estimation

Uncertainty estimation provides a measure of the uncertainty related to the model's prediction. In this work, we define an uncertainty estimation technique as a function $f_{uncertainty}$ that maps input data to a single scalar representing an uncertainty estimation value. This is typically executed on individual samples in a dataset, but can also be applied to instances in MIL. Given an instance $\mathbf{x}_i$, the uncertainty can be estimated as $U_i = f_{uncertainty}(\mathbf{x}_i)$.

Several uncertainty estimation techniques also need the selection of an uncertainty measure ($f_{measure}$) to aggregate multiple probability vectors. In the last part of this section, we also define several of these uncertainty measure functions. Some popular uncertainty estimation techniques are:

- **Deep Ensemble [230]:** This method utilizes multiple deep networks trained with various random model weight initializations. To estimate uncertainty, each model's output is considered:

$$f_{uncertainty}(\mathbf{x}_i) = f_{measure}(f_{\theta_{instance,1}}(\mathbf{x}_i), \ldots, f_{\theta_{instance,M}}(\mathbf{x}_i)),$$

  where each $f_{\theta_{instance,j}}$ is the j-th instance-level classification model in the ensemble and M is the total number of models in the ensemble.

- **Monte-Carlo dropout (MC-DO) [194]:** This method estimates uncertainty by employing dropout at test time and performing multiple forward passes:

$$\begin{aligned} f_{uncertainty}(\mathbf{x}_i) = f_{measure}(&f_{\theta_{instance,MCDO}}(\mathbf{x}_i; t_1), \ldots, \\ &f_{\theta_{instance,MCDO}}(\mathbf{x}_i; t_T)), \end{aligned}$$

  where $f_{\theta_{instance,MCDO}}$ is an instance-level classification model with dropout enabled during test-time. The term t_k represents the k-th stochastic forward pass out of a total of T passes.

- **Test-time augmentation (TTA) [247]:** This method augments the input sample at test time and observes the variation in the classifier's output. The uncertainty is cal-

culated as:

$$f_{uncertainty}(\mathbf{x}_i) = f_{measure}(f_{\theta_{instance}}(Aug_1(\mathbf{x}_i);), \ldots, f_{\theta_{instance}}(Aug_T(\mathbf{x}_i))),$$

where $f_{\theta_{instance}}$ is an instance-level classification model and $Aug_k(\mathbf{x}_i)$ represents the application of the k-th augmentation to the instance $\mathbf{x}_i$ out of a total of T augmentations.

- **MaxLogit [248]:** This technique considers the maximum logit value (before softmax) as a confidence measure. To obtain an uncertainty value, rather than a confidence value, we take the negative of the maximum logit value. The uncertainty is directly given by:

$$U_i = -\max(\mathbf{l}_i),$$

where $\mathbf{l}_i$ is the logit vector for the instance $\mathbf{x}_i$. Please note this is the only uncertainty estimation technique described in this work that does not utilize $f_{measure}$.

The list below describes a number of uncertainty measures. Each measure utilizes a function $f_{measure}$ that operates on a series of model outputs for an instance, producing an uncertainty value. Given a series of model outputs $p_{i,1}, p_{i,2}, \ldots, p_{i,T}$ for instance $\mathbf{x}_i$, where T can for example represent multiple models in an ensemble, stochastic forward passes in MC-DO, or multiple augmentations in TTA, the uncertainty measure function $f_{measure}$ produces:

$$U_i = f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}).$$

- **Maximum class probability:**

$$f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}) = -\max_{c=1}^{C} \frac{1}{T} \sum_{t=1}^{T} p_{i,t,c},$$

where U_i is the uncertainty associated with instance $\mathbf{x}_i$, C is the number of classes. $p_{i,t,c}$ is the probability assigned by the t-th model, stochastic forward pass or augmentation for instance $\mathbf{x}_i$ and class c. Since we are interested in obtaining uncertainty values, rather than confidence values, we employ a minus-sign at the start of the equation.

- **Mean class variance:**

$$f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}(p_{i,t,c} - \mu_{i,c})^2,$$

where $\mu_{i,c} = \frac{1}{T}\sum_{t=1}^{T} p_{i,t,c}$ is the mean probability of the c-th class for instance $\mathbf{x}_i$.

- **Entropy:**

$$f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}) = -\sum_{c=1}^{C}\mu_{i,c}\log\mu_{i,c},$$

where $\mu_{i,c} = \frac{1}{T}\sum_{t=1}^{T} p_{i,t,c}$ is as previously defined.

For ordinal classification tasks, such as those with a staging scale, it is critical to note that conventional uncertainty measures might not accurately capture the differences between stages. Specifically, they might treat uncertainties between neighboring stages the same as those between distant stages. Consider a scenario where probabilities for classes 1 and 2 are both 50%, rendering other classes at 0%. The uncertainty should be lower than a scenario where class 1 and class 5 have probabilities of 50%. To address this, we propose ordinal variants of mean class variance and entropy, known as ordinal variance and ordinal entropy. These measures specifically consider the ordinal relationship between classes, ensuring that larger uncertainties between distant classes are weighted more heavily than those between closer classes:

- **Ordinal variance:**

$$f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}) = \frac{1}{T}\sum_{t=1}^{T}(q_{i,t} - \mu_i)^2,$$

where $q_{i,t} = \sum_{c=1}^{C}(c-1)\cdot p_{i,t,c}$ represents the weighted sum of class probabilities by their ordinal rank for instance $\mathbf{x}_i$. This essentially converts the class probabilities to a single scalar value, similar to a value that a regular regression model would output. The term μ_i denotes the average of these weighted sums across all models or stochastic passes, calculated as $\mu_i = \frac{1}{T}\sum_{t=1}^{T} q_{i,t}$.

- **Ordinal entropy:**

$$f_{measure}(p_{i,1}, p_{i,2}, \ldots, p_{i,T}) = \\ -\sum_{c=1}^{C-1} \left(\underbrace{\sum_{d=1}^{c} \mu_{i,d} \log \sum_{d=1}^{c} \mu_{i,d}}_{\text{Entropy of classes up to } c} + \underbrace{\sum_{d=c+1}^{C} \mu_{i,d} \log \sum_{d=c+1}^{C} \mu_{i,d}}_{\text{Entropy of classes beyond } c} \right),$$

where $\mu_{i,c} = \frac{1}{T} \sum_{t=1}^{T} p_{i,t,c}$ represents the mean probability of instance $\mathbf{x}_i$ being of class c across all models or stochastic forward passes. This measure calculates multiple entropies based on binary partitions of the classes. For every class c except the last class, C, it breaks down the classes into two groups: one with classes up to c and the other with classes greater than c. By doing this, it constructs binary uncertainties for each partition and computes their entropies. These entropies are then summed up, producing a measure that captures ordinal relationships between classes.

5

5.3.3 Uncertainty-based Instance eXclusion (UBIX)

UBIX is neither a standalone MIL method, like Max pooling or Distribution pooling, nor an alternative uncertainty estimation approach, such as deep ensembles or MaxLogit. Instead, it synergistically leverages existing MIL methods and uncertainty estimation techniques. Moreover, UBIX can seamlessly integrate with a broad spectrum of MIL methods and uncertainty estimation techniques.

In UBIX, there is a deliberate inference-time manipulation of instance-level features based on the associated uncertainties. As opposed to assuming equal contribution, UBIX identifies instances that might be corrupted due to local artifacts, and reduces or ignores the contributions of these instances on-the-fly using these associated instance-level uncertainties. We can categorize two distinct UBIX variants (see Fig. 5.1), which both perform this manipulation in a different manner. Specifically, $\text{UBIX}_{\text{soft}}$ reduces the contributions of uncertain instances and $\text{UBIX}_{\text{hard}}$ fully removes their contributions.

- **$\text{UBIX}_{\text{soft}}$**: For clarity, we first present UBIX for a straightforward scenario, without the intricacies introduced by techniques such as deep ensembles or TTA. $\text{UBIX}_{\text{soft}}$ is specifically designed for scenarios where $f_{\theta_{pooling}}$ employs Max pooling in conjunction with the instance-based MIL approach. We formulate this restriction because this UBIX variant actively reduces instance-level features as uncertainty increases, a behavior that is appropriate solely when Max pooling is applied.

f_{UBIX} can be conceptualized as a modified sigmoid function. During inference, it maps the instance-level features, $\mathbf{h}_i$, based on the associated uncertainty, U_i, to updated instance-level features. These updated instance-level features are then used in the downstream MIL pipeline. The intention is for $f_{UBIX}(\mathbf{h}_i, U_i)$ to remain close to $\mathbf{h}_i$ when uncertainty U_i is minimal. Therefore, the sigmoid function approaches $\mathbf{h}_i$ as uncertainty decreases, *i.e.,* the upper asymptote is set to $\mathbf{h}_i$. Conversely, when U_i is high, we desire the logits to be diminished. Therefore, the sigmoid function converges towards the lowest feature value observed in the validation set as the uncertainty increases.

More formally, the instance-level features are transformed by the UBIX function f_{UBIX} as follows:

$$f_{UBIX}(\mathbf{h}_i, U_i) = \frac{\mathbf{h}_i - \mathbf{h}_{min}}{1 + e^{\delta(U_i - \hat{\gamma})}} + \mathbf{h}_{min},$$

where δ is a hyperparameter dictating the smoothness of f_{UBIX}. The term $\mathbf{h}_{min}$ is the vector of length J (which is equal to C in the instance-based approach), containing the minimum instance-level outputs within the validation set for each class, defined as:

$$\mathbf{h}_{min} = [\min_{\mathbf{x}_i \in \mathcal{D}_{val}} \{h_i^1\}, \ldots, \min_{\mathbf{x}_i \in \mathcal{D}_{val}} \{h_i^J\}],$$

with $\mathcal{D}_{val}$ representing the validation dataset. The steepest gradient of f_{UBIX} with respect to $\mathbf{h}_i$ occurs when U_i equals $\hat{\gamma}$, which is defined as:

$$\hat{\gamma}(\gamma) = \gamma(U_{max} - U_{min}) + U_{min},$$

with γ being a tunable hyperparameter. The extremities of uncertainties are described by:

$$U_{min} = \min_{\mathbf{x}_i \in \mathcal{D}_{val}} \{U_i\},$$

and:

$$U_{max} = \max_{\mathbf{x}_i \in \mathcal{D}_{val}} \{U_i\}.$$

In conclusion, the sole change that UBIX introduces is during the inference phase where the MIL pooling function $f_{\theta_{pooling}}$ does not directly utilize $\mathbf{h}_i$. Instead, $f_{\theta_{pooling}}$ takes the transformed instance-level features $f_{UBIX}(\mathbf{h}_i, U_i)$ as input.

When multiple forward passes are employed during inference due to the use of techniques like deep ensembles, TTA, or MC-DO, the UBIX function f_{UBIX} operates in the same way. There will be multiple values for $\mathbf{h}_i$, one for each forward pass.

$f_{\theta_{transform}}(f_{\theta_{pooling}}(f_{UBIX}(\mathbf{h}_i)))$ is performed separately for these different values for $\mathbf{h}_i$. Only the bag-level probabilities are subsequently aggregated.

- **UBIX$_{\text{hard}}$**: This variant of UBIX operates by fully excluding certain instances based on their uncertainty values. Specifically, the UBIX function is defined differently in this approach:

$$f_{UBIX}(\mathbf{h}_i, U_i) = \begin{cases} \mathbf{h}_i & \text{if } U_i \leq \tau \\ \text{exclude} & \text{otherwise} \end{cases}$$

where τ is a hyperparameter that determines the uncertainty threshold for instance exclusion. In this approach, any instance with an uncertainty value exceeding τ is completely excluded from further processing.

When implementing UBIX$_{\text{hard}}$, the MIL pooling function, $f_{\theta_{pooling}}$, is fed with a subset of instance-level features after exclusion.

5.4 Data

5.4.1 Data

Three different data sets from three different vendors were used to develop and evaluate the proposed solution: a dataset with Heidelberg OCTs served as a training set, referred to as H_{train}, a validation set, referred to as H_{val}, and internal test set, referred to as H_{test}; and two external test sets were used to evaluate the generalizability of our models, one with

Table 5.1: CIRCL grading system.

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

Table 5.2: AMD stage distribution in each dataset. The table shows the number of OCT scans.

	H_{train}	H_{val}	H_{test}	T_{test}	B_{test}
No AMD	597	216	212	942	115
Early AMD	202	74	70		
Intermediate AMD	329	106	114	149	269
Advanced AMD: GA	72	29	28	37	
Advanced AMD: CNV	737	236	256	56	

Topcon scans, referred to as T_{test}, and one with Bioptigen scans, referred to as B_{test}.

Heidelberg dataset (H_{train}, H_{val} and H_{test})

For development and internal testing, we used the European Genetic Database (EUGENDA), a large multi-center database for clinical and molecular analysis of AMD [249, 250], containing 3,278 OCT from 1,013 patients in total. The training set H_{train}, validation set H_{val}, and test set H_{test} contained 1937, 661 and 680 OCTs from 607 (60%), 202 (20%) and 204 (20%) patients, respectively.

Manual grading of the scans was performed by the Cologne Image Reading Center and Laboratory (CIRCL). They categorized the OCTs using the criteria described in Table 5.1. Samples with grade 6 and 7 were excluded from this study. The number of OCTs for each of the remaining five stages is shown in Table 5.2. The OCTs were acquired with a Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany) scanner. We resampled all B-scans to the same pixel spacing of 13.9 µm × 3.9 µm. The number of B-scans in each OCT scan was left unchanged, which varied from 14 to 73.

Topcon dataset (T_{test})

One of the external test sets was derived from the Rotterdam Study [251]. This is a prospective cohort study in the city of Rotterdam, the Netherlands, that started in 1990 to investigate age-related diseases. There were in total 1184 OCT scans available from this dataset, originating from 713 patients. All OCTs were graded using the Wisconsin age-related maculopathy grading system (WARMGS) [252] and manually harmonized to the CIRCL grading system. The number of OCTs for the resulting four classes is shown in Table 5.2. The OCTs from this dataset were taken with an OCT scanner from Topcon Corp., Tokyo, Japan. Each OCT volume contained 128 B-scans. Similarly to the Heidelberg set, all B-scans were resampled to have a pixel spacing of 13.9 µm × 3.9 µm.

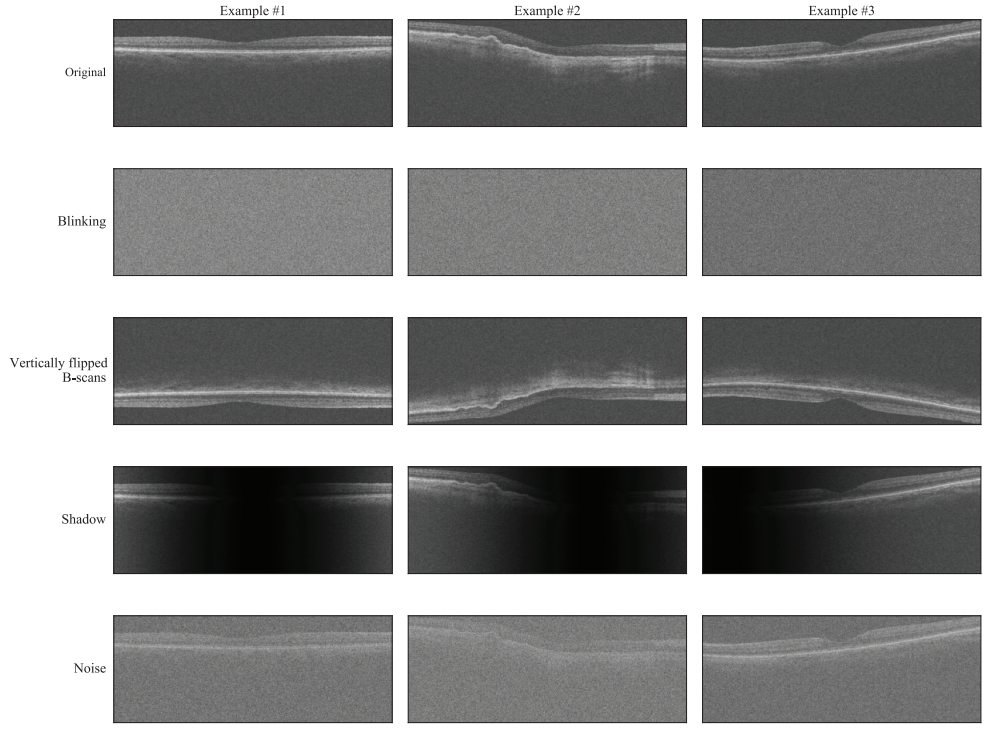


Figure 5.3: Examples of artificial artifacts. Each row shows the middle B-scans of a random OCT volume from T_{test}. The image in the first row is the original, unaltered B-scan. The other rows each depict a different artificial artifact applied to that B-scan.

Bioptigen dataset (B_{test})

The other external test set was described by [253], containing normal patients and patients with intermediate AMD. For each of these subjects, one OCT volume, acquired with an SD-OCT scanner from Bioptigen, Inc (Research Triangle Parc, NC), was available. The AREDS2 system [254] was used for grading and was harmonized to the CIRCL grading system. The number of OCTs for these two classes is given in Table 5.2. All OCT volumes contained 100 B-scans and all B-scans were again resampled to have a pixel spacing of 13.9 µm $\times$ 3.9 µm.

5.5 Experimental design

5.5.1 Vendor generalizability and interpretability

To assess the vendor generalizability of UBIX, we first calculated the performance on the internal test set, H_{test}, which is from the same distribution as the one used for training,

H_{train}. Subsequently, we evaluated the performance when transferring to the two external datasets T_{test} and B_{test}. We evaluated the performance on these datasets for $UBIX_{soft}$ and $UBIX_{hard}$. Additionally, we compared the performance of the proposed model with three different approaches, namely a 3D CNN approach, a traditional MIL approach (without UBIX) and an ensemble of multiple MIL approaches. The 3D CNN was a ResNet-18 [255] with 3D convolutions and the instance-level classifiers in the MIL approaches were ResNet-18's with 2D convolutions. All models in our experiments were trained end-to-end.

To better show the effect of the proposed methodologies on scans with vendor-specific artifacts, we also separately evaluated the performance of the five aforementioned UBIX variants and baselines on a subset of OCT volumes in T_{test} with blinking artifacts, referred to as T_{blink} (n = 33). These volumes generally have multiple B-scans in which the retina is not visible.

The interpretability of UBIX is illustrated qualitatively by showing the instance-level predictions and uncertainties for several OCT scans.

5.5.2 Effect of artificial artifacts

To demonstrate the effect of UBIX more clearly, we performed experiments where we artificially corrupted the dataset T_{test} with artifacts that also occur naturally in OCT scans. The different artifact types were blinking artifacts, vertically flipped B-scans, shadows and noise. Fig. 5.3 shows a number of examples. We gradually introduced more OCT volumes with artificial artifacts and compared the performance for $UBIX_{soft}$, $UBIX_{hard}$ and MIL.

When one of these artifacts was applied to an OCT volume, a portion of the B-scans were affected, as happens in clinical scenarios. Artificial artifacts were then added to either one or two groups of adjacent B-scans. Both scenarios had an equal probability. The sizes of these groups had sizes of between 2% and 15% of B-scans, which we experimentally found to be representative of real artifacts.

Vertically flipped B-scans are caused by a Fourier-domain detection artifact, as described by [256]. Shadows and noise are usually caused by media opacities, such as corneal scarring and cataract. The artificial artifacts were implemented as follows:

- To generate B-scans with artificial blinking artifacts, we started with an image where all pixel values were set to 0. Next, we added random Gaussian noise to this image. The mean of this noise was set to the median pixel value found in the full OCT scan. The standard deviation of this noise was matched to the standard deviation of the OCT scan.

- The vertically flipped B-scans were generated by flipping the B-scans along the horizontal axis.

- To generate the shadow artifact for a particular B-scan, we adapted each A-scan (column in an OCT B-scan) $a_1, \ldots, a_A$ separately, where A is the number of A-scans in the B-scan. All A-scans a_i were transformed using the shadow function $S(a_i)$:

$$S(a_i) = a_i(1 - s(i)), \tag{5.1}$$

where $s(i)$ is sampled from a normal probability density function:

$$s(i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{i-\mu}{\sigma})^2}, \tag{5.2}$$

where μ is randomly selected between 0 and A, and σ is randomly defined between $A/4$ and $3A/4$. μ and σ are kept the same within one OCT volume.

- The noise artifact is Gaussian noise added to the original image with a mean of 0 and a standard deviation of 4 times the standard deviation within the original OCT volume.

5.5.3 Applicability to several MIL settings and uncertainty estimation approaches

We performed several experiments to assess the applicability of UBIX to different uncertainty estimation approaches and MIL settings. We evaluated this on the external test sets T_{test}, T_{blink}, and B_{test}. The uncertainty measures and techniques that we evaluated were the ones presented in Section 5.3.2. The MIL pooling functions we used in these experiments were described in Section 5.3.1.

Some of the MIL pooling functions are only compatible with the embedding-based approach. Therefore, in the experiment where we compared these MIL pooling functions, all uncertainties were calculated using MaxLogit, considering the other uncertainty measures cannot be combined trivially with the embedding-based approach. Max and Mean pooling were implemented using the instance-based MIL approach, while the Attention, Distribution, Distribution with attention, and TransMIL pooling were implemented using the embedding-based approach. For Attention pooling, we followed the implementation provided by [212]. For Distribution and Distribution with Attention pooling, we used the implementation provided by [244] with 128 features in the attention layers and 11 bins in the distribution pooling layer. For TransMIL pooling, we used the implementation provided by [246]. The deep ensembles contained 5 models, as this was shown to be sufficient for

uncertainty estimation [208]. For the MC-DO models, we used $T = 32$ stochastic forward passes, which is consistent with what was done by [215]. For TTA we used $T = 10$ different forward passes. For the experiments with MaxLogit, we used the maximum value in the output of the final layer of the instance-level classifier, while only considering this maximum value for the first model in case a deep ensemble was used. The UBIX hyperparameters were separately optimized on the validation set H_{val} for each unique combination of MIL pooling function and uncertainty estimation approach.

5.5.4 Metrics

For all models, to evaluate the classification performance, we calculated the area under the receiver operating characteristic curve (AUC), where intermediate and advanced AMD stages belonged to the positive class and the remaining stages to the negative class. Additionally, we computed Cohen's kappa score. For the datasets with more than two classes, the quadratic weighted kappa score (κ_w) was calculated to consider the class order. Otherwise, unweighted kappa metric was used (κ).

To quantify artificial artifact detection performance for different uncertainty measures, we used the AUC as well, where the score was the uncertainty measure and the labels were the dichotomous variable of whether an instance had an artificial artifact or not. Furthermore, to estimate how well the uncertainty values were separated, we evaluated the separability of the two groups, with and without artificial artifacts, based on the uncertainty score. For this, the Xie-Beni index (XB) is calculated, defined as the ratio between cluster separation (*i.e.*, the minimum squared distance between cluster centers) and cluster compactness (*i.e.*, the mean squared distance between each data point and its cluster center [257]. The lower XB, the better the data is clustered. Statistical significance was determined using non-parametric bootstrapping with 1000 iterations [127]. We applied a Bonferroni correction to account for the number of comparisons we made.

5.5.5 Training and optimization

The network weights were optimized with the Adam optimizer [258] and a learning rate of 10^{-4} using the cross entropy loss. All images were normalized between 0 and 1. As a means of regularization, we employed online data augmentation. With a 15% probability, random affine transformations were applied of ±20° rotation within the B-scan plane, ±10% shearing within the B-scan plane, ±10% zooming within the B-scan plane, ±20 voxels translation in the horizontal and vertical direction within the B-scan plane, ±2 voxels translation in the B-scan direction. B-scans were also horizontally flipped with a 15% probability. With a 30%

probability, random additive Gaussian noise with a mean of 0 and a standard deviation of 0.1 was applied. Also with a 30% probability, we applied brightness modifications using the power law, varying the power between 0.75 and 3. To account for class imbalance, images were sampled during training based on their class, such that each class was sampled with an equal probability. Because of GPU memory restrictions and given that the input images were large, each batch contained one bag during training and inference. We used early stopping, based on κ_w and with a patience of 10 000 batches.

The hyperparameters τ, δ and γ were optimized on the internal validation set $\mathrm{H}_{\mathrm{val}}$. The values for τ in the grid search for $\mathrm{UBIX}_{\mathrm{hard}}$ were $\{P_{U,80}, P_{U,80.1}, \ldots, P_{U,100}\}$, where $P_{U,i}$ is the percentile i of all uncertainties in the validation set. For the UBIX grid search, the values for δ were $\{1, 5, 10, \ldots, 5 \cdot 10^3, 10^4\}$ and the values for γ were $\{-0.5, -0.45, \ldots, 1.5\}$. For $\mathrm{UBIX}_{\mathrm{soft}}$ implemented using deep ensembles, ordinal entropy and Max pooling, the performance was optimal at $\delta = 5$ and $\gamma = 1.05$. For $\mathrm{UBIX}_{\mathrm{hard}}$ implemented using deep ensembles, ordinal entropy and Max pooling, the optimal value of the hyperparameter τ was $P_{U,93.3} = 2.20$.

5.5.6 Normalization type

All MIL and UBIX models (which do not differ from MIL models during training) described in this section so far used batch normalization (BatchNorm) [259] followed by a rectified linear unit (ReLU) activation layer [260] after every convolutional layer. This BatchNorm layer was implemented such that the mean and variance were computed along the instance axis (rather than the bag axis), spatial, and channel axes. This had the effect that the instance-level classifier in these models used some information from other instances than the instance it was classifying, specifically when calculating the mean and variance values that are used in this layer. We studied the effect of replacing these BatchNorm layers with InstanceNorm [261] layers on the generalization of both the MIL and UBIX models. For InstanceNorm, the mean and variance values were only calculated along the spatial and channel axes, so not along the bag or instance axes. Following the default implementation of PyTorch, the BatchNorm models included learnable shift and scale parameters, while the InstanceNorm models did not. The 3D model used in this paper also had a batch size of 1, so the normalization layer was effectively InstanceNorm with learnable shift and scale parameters. In all cases, the means and variances of the current sample, rather than the moving averages, were used during both training and inference.

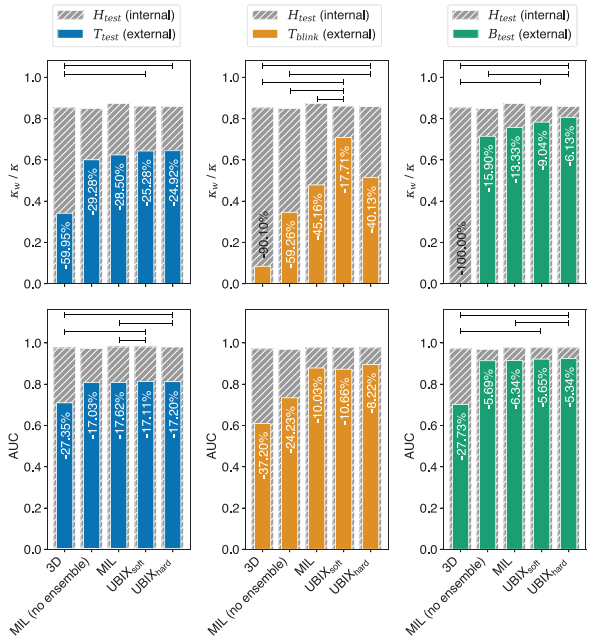


Figure 5.4: Performance of UBIX, compared to the 3D and MIL baselines, evaluated on internal and external datasets. The models were all trained and validated on Heidelberg data (H_{train} and H_{val}). The UBIX methods in this figure use deep ensembles with ordinal entropy to estimate uncertainty. The percentages displayed within the bars indicate the difference in performance when transferring the model from the internal to the external test set. Max pooling was used for each of the reported results. The horizontal black bars above the bar plots indicate whether there was a statistically significant difference ($p < 0.05$) between the performance on the external datasets of either of the UBIX variants and the other three models.

5.6 Results

5.6.1 Baseline comparison

In this section, we compare the performance of the two UBIX variants with three baseline models on the internal dataset. Additionally, we explore how these models differ in performance when applied to external datasets. Fig. 5.4 provides an overview of these comparisons.

On the internal dataset, we observe that most differences between models are relatively small compared to the external datasets. When evaluating on external datasets, there is a notable drop in performance for all methods. However, this drop is consistently smaller for the UBIX models when considering κ_w as the performance metric.

UBIX exhibits superior performance in scenarios characterized by the presence of artifacts. For instance, on the T_{blink} dataset, UBIX achieves a κ_w of 0.708, surpassing the scores

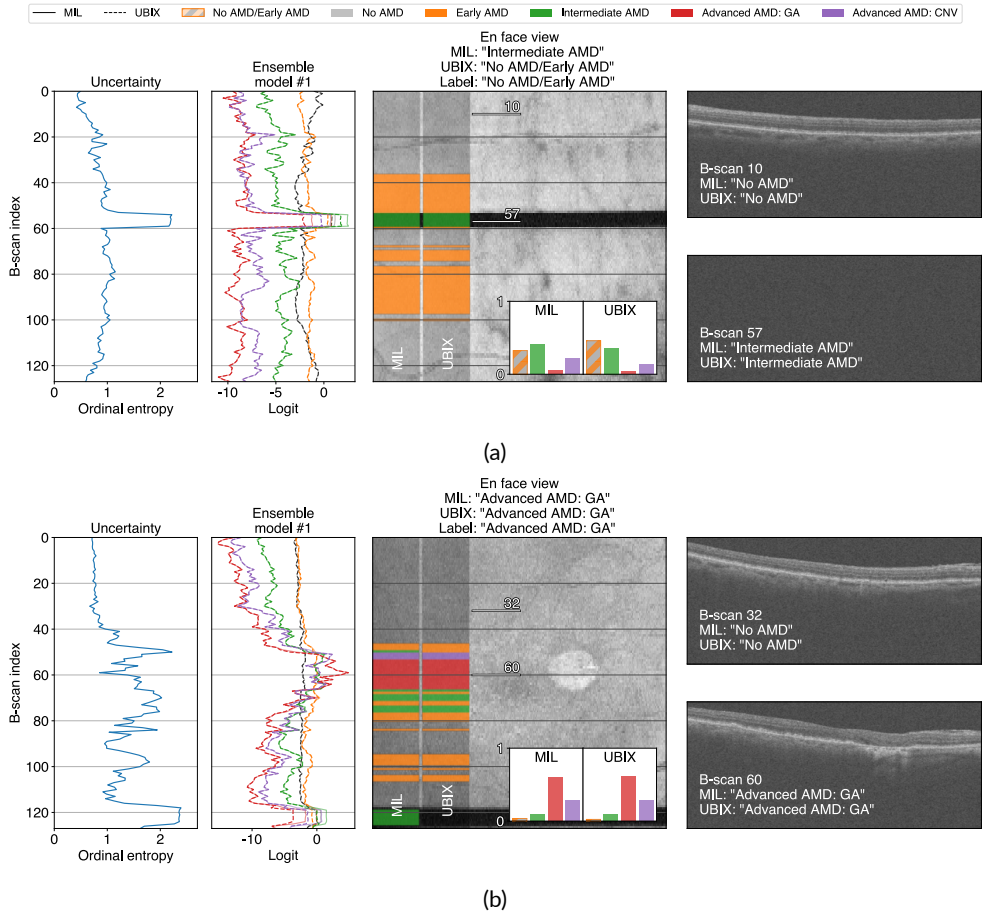


Figure 5.5: Examples where $\text{UBIX}_{\text{soft}}$ corrects volume-level and instance-level predictions. The figure also illustrates instance-level interpretability. Each subfigure shows, from left to right, the uncertainty per instance, the instance-level logits of the first model in the ensemble (only showing one model for clarity), the *en face* image (the volume averaged over the y-axis), and two B-scans of interest. The uncertainty, logit and *en face* plots correspond spatially in the horizontal direction. The left and right banner in the *en face* view indicate the instance-level outputs for MIL and UBIX, respectively. The B-scans on the right are highlighted in the *en face* view. In the bottom right of the *en face* views, volume-level probabilities are shown. **(a)** UBIX correctly predicts "No AMD/Early AMD", while MIL incorrectly predicts "Intermediate AMD". UBIX suppresses the instance-level outputs at the location of the artifact around B-scan 57, causing it to be robust to that artifact, in contrast to MIL. **(b)** Correct GA volume-level classification. Central GA is also well visible in the *en face* image as a white circular object around B-scan 60, which is pointed out by the instance-level output in red.

of 0.479 for only MIL, 0.346 for MIL (no ensemble), and 0.084 for the 3D model. To illustrate the interpretability of the model at the B-scan level, we provide visual examples of $\text{UBIX}_{\text{soft}}$ predictions in Fig. 5.5.

In Fig. 5.6, we investigate the impact of introducing artificial artifacts when using UBIX, compared to MIL. We evaluate performance while varying the percentage of OCT volumes

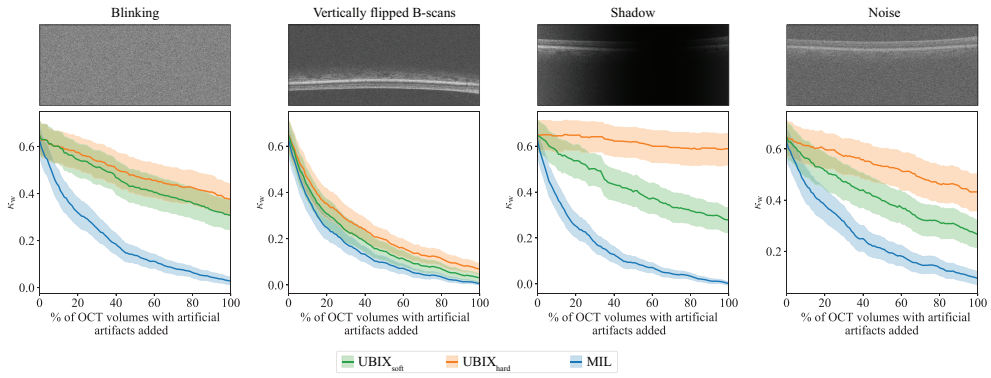


Figure 5.6: Robustness to different artificial artifacts of UBIX, compared to MIL, on T_{test}. The top image in each column shows an example B-scan of the artificial artifact. The plots show the relation between the performance and the percentage of OCT volumes in the dataset that contain these artifacts. The shaded areas indicate 95% confidence intervals, obtained using bootstrapping with 1000 iterations.

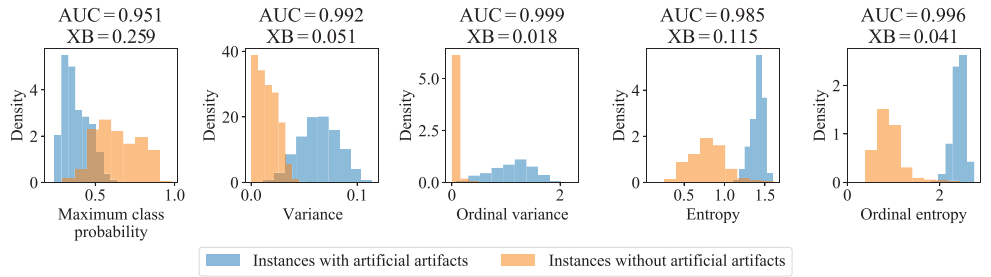


Figure 5.7: Density plots of uncertainty estimates at instances with artificial blinking artifacts compared to uncertainty estimates at instances without artificial artifacts, applied to T_{test}. The AUC for artificial artifact detection performance and the clustering metric XB (the lower, the better) are shown above the density plots.

affected by artificial artifacts within the T_{test} dataset.

5.6.2 Applicability to several MIL settings and uncertainty estimation approaches

We explored the applicability of $UBIX_{soft}$ in different MIL settings and uncertainty estimation approaches. Fig. 5.7 and 5.8 show the generalizability of $UBIX_{soft}$ when using different uncertainty measures. The latter figure shows that for T_{test} and κ_w, ordinal entropy emerges as the top-performing measure. For B_{test}, the two ordinal variants also show the highest performance.

Several uncertainty estimation methods were compared in Fig. 5.9. For T_{blink}, the biggest impact of UBIX could be observed when using deep ensembles as the uncertainty estimation method. Fig. 5.10 shows the generalizability of several MIL pooling functions and the effect of UBIX in combination with these different MIL pooling functions.

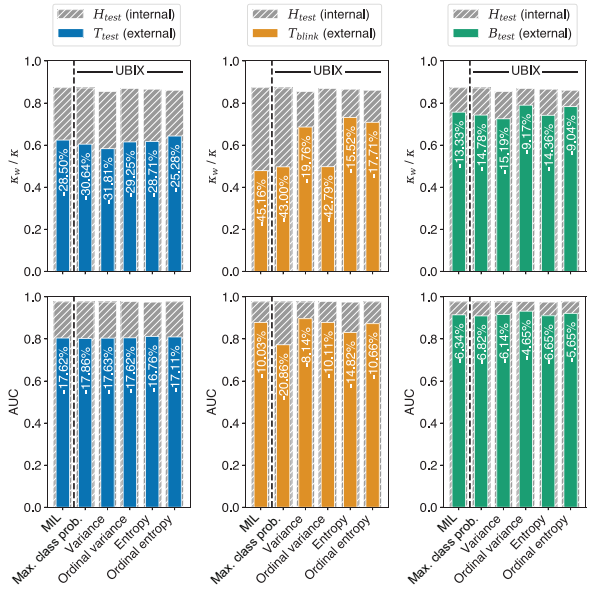


Figure 5.8: Performance metrics for different uncertainty measures used for UBIX$_{soft}$, evaluated on the internal and external test sets. The MIL method performance (without UBIX) is indicated on the left of each plot. The models were all trained and validated on Heidelberg data (H_{train} and H_{val}). Deep ensembles were used in combination with the reported uncertainty measures. Max pooling was used for each of the reported results. Max. class prob. = Maximum class probability.

5.6.3 Normalization type

The effect of using InstanceNorm instead of BatchNorm on generalizability for MIL and UBIX is shown in Fig. 5.11. Since the drop in performance is already substantially less when transferring to the external datasets for the MIL models with InstanceNorm than with BatchNorm, the effect of UBIX is less visible in this figure. However, when considering images with different types of artifacts than blinking artifacts, we found that UBIX was able to recover better than MIL, as illustrated in Fig. 5.12.

5.7 Discussion

We proposed a method using MIL with OOD detection to improve the generalizability of deep learning models for classification of 3D medical images. The model aims to reduce the effect of on-the-fly detected OOD instances in the final classification of the bag. By suppressing the contribution of OOD instances, UBIX maintains performance on unseen data distributions, particularly images coming from different scanners.

The robustness of the proposed approach was demonstrated by transferring UBIX mod-

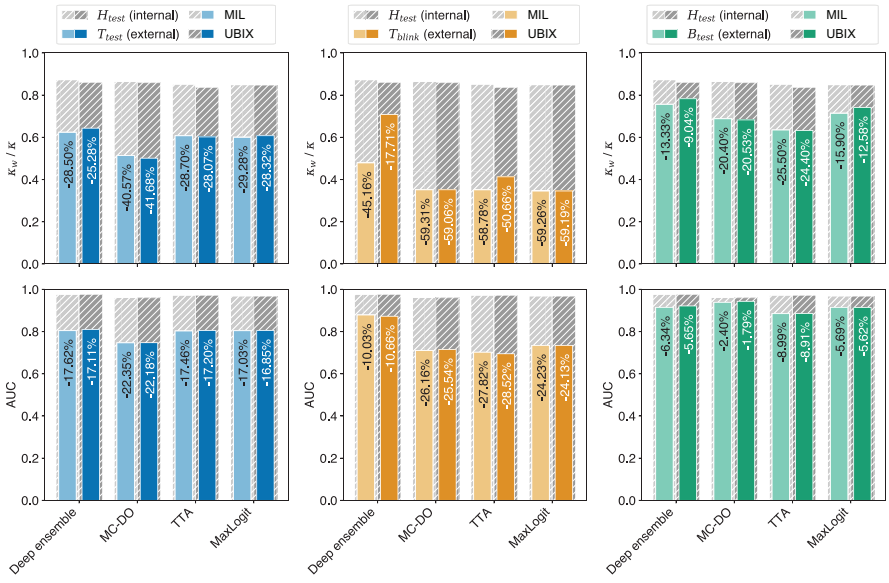


Figure 5.9: Performance metrics for different uncertainty estimation methods used for $UBIX_{soft}$, compared to MIL (without UBIX), evaluated on the internal and external test sets. The models were all trained and validated on Heidelberg data (H_{train} and H_{val}). Max pooling was used for each of the reported results. For the MC-DO, TTA, and MaxLogit experiments, one model was used, instead of an ensemble.

els and baseline models to external datasets from different vendors. As shown in Fig. 5.4, UBIX variants are less prone to substantial performance drops than the other models. On all external datasets, either $UBIX_{soft}$ or $UBIX_{hard}$ showed better results than the baseline models in terms of absolute performance. The performance drop was most notable on T_{blink}, where $UBIX_{soft}$ maintained a κ_w of 0.708, while the best and worst performing baseline models (MIL and 3D, respectively) had a κ_w of 0.479 and κ_w of 0.084, respectively. It was expected that these performance differences were more notable on T_{blink}, which only contained OCTs with blinking artifacts, because UBIX was designed to be robust to vendor-specific artifacts.

It was noted that, depending on the inference data, it was preferable to fully exclude the outputs of uncertain instances ($UBIX_{hard}$) or to only suppress them ($UBIX_{soft}$). From Fig. 5.6, we found that $UBIX_{hard}$ showed better robustness than $UBIX_{soft}$ for OCTs with artificial artifacts. A possible reason for this could be that some of the artificial artifacts highly corrupted the information, resulting in a notably strong incorrect signal and the requirement for full exclusion of uncertain instances. $UBIX_{hard}$ seemed to be especially robust to shadow artifacts, given that the performance barely decreased when introducing more OCT volumes with artificial artifacts. For some external datasets and metrics, however, $UBIX_{soft}$ achieved better results than $UBIX_{hard}$, *e.g.*, for κ_w on T_{blink} and AUC on T_{test}.

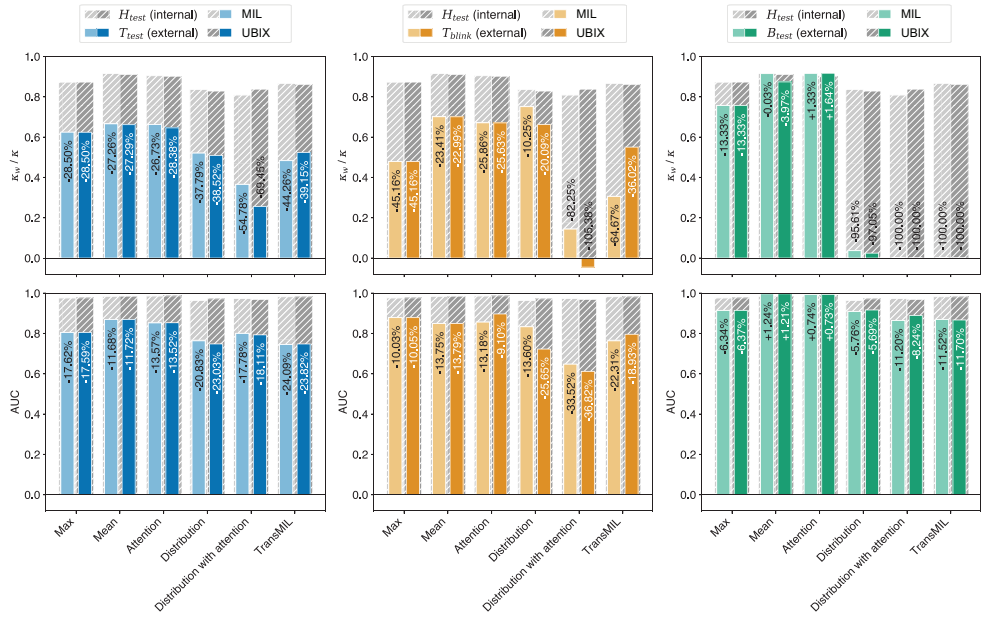


Figure 5.10: Performance metrics for different MIL pooling functions when using $UBIX_{hard}$, compared to MIL (without UBIX), evaluated on the internal and external test sets. The models were all trained and validated on Heidelberg data (H_{train} and H_{val}). $UBIX_{hard}$ was used because $UBIX_{soft}$ was only designed for MIL models that utilize Max pooling. Uncertainties were calculated using MaxLogit in all reported UBIX models for this experiment. Ensembles of models were used for prediction in this experiment.

The performed data augmentation might also have an effect on generalizability. Since signal-to-noise ratios differ per scanner, noise augmentation probably aided our models to generalize. Although we did resample all images to have the same pixel spacing within B-scans, the original spacings that we had could have been slightly inaccurate. Therefore, zooming could potentially also have a positive effect on generalizability. The same type of data augmentation was applied in all experiments and measuring its effect was considered out of the scope for this paper.

Large variability in B-scan spacing between scanners can also cause features learned by 3D CNNs to be poorly generalizable. MIL, which processes B-scans individually and combines B-scan level outputs using a MIL pooling function to get a volume level output, improves the robustness to this variability in slice spacing with respect to 3D models. We observed that performance differences are minimal between a 3D CNN and MIL when evaluated on data from the same vendor used during training. When evaluating on data from a different vendor, a performance drop was observed for both methods, although this drop is much larger for the 3D method (60.0% drop in κ_w on T_{test}) than for MIL (29.3% in κ_w on T_{test}).

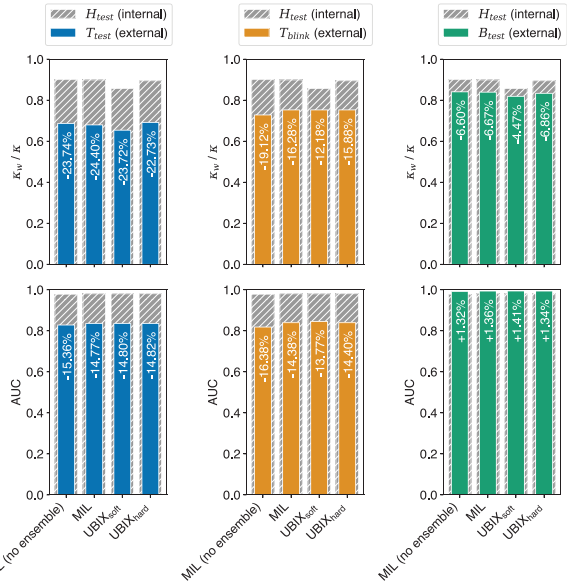


Figure 5.11: Performance of models using InstanceNorm instead of BatchNorm. Results are shown for UBIX, compared to the MIL alternatives without UBIX, evaluated on internal and external datasets. The models were all trained and validated on Heidelberg data (H_{train} and H_{val}). The UBIX methods in this figure use deep ensembles with ordinal entropy to estimate uncertainty and Max pooling as a MIL pooling function. The percentages displayed within the bars indicate the difference in performance when transferring the model from the internal to the external test set.

One of the advantages of using MIL as base of the UBIX model is that instance-level annotations are not required for training, while the model is able to produce a classification output at this level (B-scan level in our case) as well as calculating instance-level uncertainty. This introduces model explainability, increasing the transparency of our method, and allowing surveillance of its behavior.

The extent to which UBIX improves generalizability depends highly on the quality of its underlying OOD detection. Therefore, we compared three different commonly used uncertainty measures in combination with deep ensembles, and we proposed two ordinal variants. On T_{test}, entropy and its ordinal variant had the highest performance in terms of κ_w and AUC, respectively. The ordinal variants seemed to distinguish the B-scans with and without artificial artifacts the best. This can be seen in the density plots of Fig. 5.7, and this is also reflected in the AUC and XB values, which were best for the two ordinal variants. Hence, the ordinal variants led to higher performances on artificial artifacts, although in the external test sets, we find mixed observations (see Fig. 5.8). A possible reason for this could be that fewer evaluation data with real artifacts were available, resulting in a less accurate performance measurement than when using artificial artifacts.

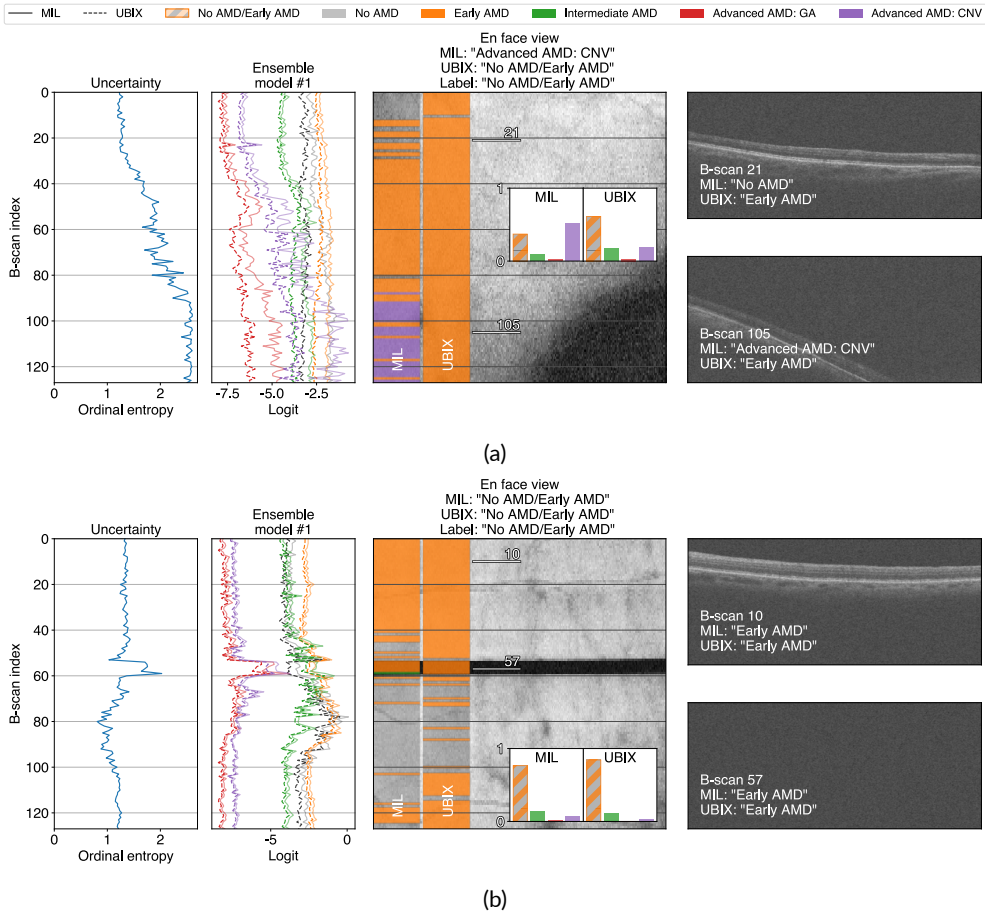


Figure 5.12: Examples of $\text{UBIX}_{\text{soft}}$ predictions when using InstanceNorm instead of BatchNorm. The subfigures follow the same structure as Fig. 5.5. **(a)** A case where B-scans corresponding to the lower part in the *en face* view suffer from an artifact that causes the retina to be displayed diagonally, falling out of view on the right. This artifact was likely out-of-distribution, resulting in the classification of these B-scans as "Advanced AMD: CNV", while they actually do not show any signs of neovascularization. The ordinal entropy was also high for these cases with artifacts, allowing UBIX to correct the bag-level prediction. **(b)** The same case was displayed in Fig. 5.5a. As could be observed in that figure, the blinking artifact caused an incorrect classification of the MIL model with BatchNorm, which was corrected by UBIX. However, the MIL model with InstanceNorm, as shown in the current subfigure, already showed robustness to this artifact without UBIX.

Furthermore, uncertainty measures for which we found the most competitive performance in terms of κ_w did not always perform the best in terms of AUC. A possible explanation for this inconsistency could be that the properties these two metrics measure are quite different. κ_w measures the grading performance (*i.e.*, the capability of distinguishing – in case of T_{test} and T_{blink} – four different grades), while the AUC only measures binary classification performance (*i.e.*, determining whether the sample has a grade higher or lower than

Intermediate AMD). Therefore, we believe that for models with varying results like these, it should be carefully considered which metrics are most relevant in which clinical settings. In screening scenarios, for example, this binary classification task (using the AUC as metric) may be more relevant. Conversely, in secondary care, a more granular grading (using κ_w as metric) may be desired.

In addition to the integration of various uncertainty measures with UBIX, we also examined its performance with uncertainty estimation techniques other than deep ensembles. As depicted in Fig. 5.9, deep ensembles generally demonstrated the least performance degradation when assessed on external datasets. This was particularly evident in the subset featuring only blinking artifacts, denoted as T_{blink}. Notably, the effectiveness of UBIX was most pronounced when paired with deep ensembles. This suggests that deep ensembles may yield more precise uncertainty estimations compared to the other three methods. This would allow UBIX to more effectively diminish the logits for instances in which this is needed (*i.e.,* where the model is prone to incorrect outputs due to OOD characteristics). This aligns with findings in the literature [208], where deep ensembles were identified as superior in detecting data shifts, outperforming other OOD detection techniques.

The observed variation in internal performance among different MIL pooling functions was relatively small, particularly compared to their external performance (see Fig. 5.10). In our comparative study of these MIL pooling functions, incorporating UBIX showed minimal effect. This minimal effect is likely attributable to our choice of MaxLogit for uncertainty estimation, a method identified as having limited influence (Fig. 5.9). Since not all MIL pooling functions could be implemented with the instance-based approach, we were restricted to MaxLogit in this experiment. We leave it to future research to explore other uncertainty estimation techniques that may be more compatible with embedding-based MIL pooling functions. In conjunction with Attention pooling on the T_{blink} dataset, adding UBIX did show a small increase in performance. Both Mean and Attention pooling outperformed others on the internal test set and the external B_{test} set. The unexpected efficacy of Mean pooling, despite initial assumptions about its inadequacy in aggregating instance scores [212], might be attributed to the nature of the reference standard, which relies on counting drusen across multiple instances in the bag. Mean pooling may be able to do this more effectively than Max pooling. The Distribution and Distribution with attention pooling functions showed no substantial inferiority compared to the other functions on the internal test set. However, on external test sets, they demonstrated underperformance in terms of κ_w. Interestingly, these effects were less pronounced for the AUC than for κ_w, suggesting a consistent shift in underlying probabilities.

Given the current experimental setup, it is challenging to recommend a single general

strategy for future practice regarding the best combination of uncertainty estimation technique and MIL pooling function for UBIX. In our experiments, we found that the optimal combination depended on multiple factors, such as whether these approaches were used with $\text{UBIX}_{\text{soft}}$ or $\text{UBIX}_{\text{hard}}$, the evaluation dataset, and the metric of interest. However, based on the experiments presented in Fig. 9 and Fig. 10, certain combinations frequently performed best across various datasets and metrics. Specifically, for $\text{UBIX}_{\text{soft}}$ with Max pooling, deep ensembles as the uncertainty estimation technique came out best most often. Furthermore, for $\text{UBIX}_{\text{hard}}$ with deep ensembles, Mean or Attention as the MIL pooling function frequently yielded the best results. Therefore, we expect these should generally be good options, even though more research is needed to provide more general recommendations.

As discussed in Section 5.5.6, the BatchNorm instance-level classifiers used in this work may use information from multiple instances, not just the ones currently being classified. We hypothesized that this characteristic might negatively impact model robustness. To test this, we also trained models using InstanceNorm, which does not share this characteristic. We found that replacing BatchNorm with InstanceNorm already improved the generalizability substantially (see Fig. 5.11). Because this generalizability without UBIX was already high, the effect of adding UBIX on the dataset-level metrics was negligible. However, even with the InstanceNorm model, we found rare individual imaging artifacts in T_{test}, to which the default model was not robust. An example of such artifacts is shown in Fig. 5.12a. We observed that with the addition of UBIX, the model was able to recover. Moreover, as shown before in this section, we showed that UBIX, for certain model choices, did enhance robustness to OOD artifacts in terms of dataset-level metrics. This suggests that UBIX is beneficial in cases of low generalizability and neutral otherwise.

Since the three datasets were acquired and annotated at different sites with varying protocols, the reference standards were set differently among these datasets. To minimize the effect of this discrepancy, we merged the first two classes of the CIRCL grading when evaluating with the WARMGS system which was available for T_{test}. Moreover, when evaluating on B_{test} for which only the binary labels *No AMD* and *Intermediate AMD* were available, we also binarized the CIRCL systems, where the positive class started at *Early AMD*. This harmonization approach, however, was not perfect, causing the resulting class definitions to still not be completely equal. Despite this discrepancy, we think measuring performance differences between methods is still well possible. Nevertheless, the absolute performances can be underestimated because of these differences in reference standards.

As a potential undesired side effect, it should be noted that difficult cases, which are assumed to be more uncertain, could be excluded which are in fact necessary for mak-

ing a correct prediction. It will depend on the setting of model deployment whether the benefits of robustness to OOD data outweigh this drawback. In screening, for example, a high specificity is generally considered more important than a high sensitivity. Instances that are falsely excluded by UBIX will often contain abnormalities, in which case especially sensitivity would suffer, but specificity will not be affected in that case. When the task is disease staging (as is the case in this work) and UBIX is implemented with an ordinal uncertainty measure, this potential undesired side-effect is unlikely to occur. Difficulty in such a case will usually lie in the uncertainty between two classes that are close together in the staging scale (such as *Early AMD* and *Intermediate AMD*), resulting in a low estimate of the ordinal uncertainty measure. As artifacts are not likely to cause uncertainties between two classes that are close on the staging scale, these artifacts will probably have a much higher ordinal uncertainty measure. To give an indication of which instances were assigned the highest uncertainties, we manually analyzed the B-scans with the highest ordinal entropies in T_{test} in Section 5.A of the Appendix. There we found that more than half of the B-scans in the first percentile of most uncertain ones contained one of the nine different types of artifacts that were seemingly related to image acquisition.

If relevant structures are entirely excluded because they are difficult for the model to classify (for example because of an unseen lesion type or ambiguity), this is likely to be because there is something atypical with the whole scan or setting in which the model is used and the user should be alerted. So instead of only silently excluding instances, future work could analyze a method for combining UBIX with alerting the user if there are too many OOD instances detected. In future work, UBIX could also be adapted to work with patches as instances instead of slices. If an artifact is only locally present within a slice, the entire slice would not be excluded and potentially useful information would not be ignored.

We did not compare our method to any other domain adaptation methods which often require additional supervised or unsupervised training. Such a comparison would be unfair, as our method does not require any additional training. Nevertheless, our approach could potentially be further improved with the incorporation of domain adaptation methods such as those proposed by [242] and [243].

Future work could also investigate different OOD detection methods to be incorporated in UBIX, as UBIX is theoretically compatible with any OOD detection method. Performing a systematic analysis for OOD detection methods was beyond the scope of this paper, so we only applied deep ensembles in this study. Such a comparison might lead to valuable insights and performance improvements.

We only evaluated our approach for AMD grading in OCT. The method is expected to be applicable in more problem settings, such as the classification of other features and retinal

diseases in OCT, but potentially also in other medical image analysis applications. Potential applications include histopathology, in which MIL is already common practice [212, 262] and many types of artifacts can occur [216,263,264], and CT scans, which can contain metal artifacts [217].

5.8 Conclusion

We showed that the generalizability of classification models to unseen scenarios can be improved by UBIX, an approach that seamlessly suppresses the contribution of OOD instances to the final classification during test-time based on the uncertainty associated with these instances, in the context of MIL for AMD classification in OCT. Our proposed approach alleviates the need for retraining on new data, which is an expensive process in terms of data acquisition, model development, and human annotation time. This increases reliability by improving the applicability of artificial intelligence models for OCT classification in broader scopes than the settings in which they were developed.

Acknowledgements

This research was funded by Eurostars grant E12712.

Chapter appendix

5.A Analysis of most uncertain B-scans

In this section, we analyze the most uncertain B-scans in the Topcon (external) test set, of which the contributions to volume-level outputs were reduced the most.

The analysis was done on the first percentile of B-scans with the highest ordinal entropy (n = 1,516). We subsequently considered the embeddings (the latent features before the fully connected layer in the first 2D ResNet-18 model of the MIL ensemble) of these B-scans and reduced the dimensionality using t-SNE ([265]). This resulted in the plot in Fig. 5.A.1. We also looked at this first percentile of most uncertain B-scans and reported for those B-scans if we found any artifact or abnormality that was seemingly related to image acquisition. In total we found 9 different types of such artifacts (see the embeddings and examples per artifact type in Fig. 5.A.2, 5.A.3, 5.A.4, 5.A.5, 5.A.6, 5.A.7, 5.A.8, 5.A.9 and 5.A.10). Embeddings and examples of B-scans without any abnormalities related to image acquisition can be found in 5.A.11.

t-SNE does not perfectly separate all classes, possibly because the model was trained to separate AMD stages. Hence, its features were mainly optimized for that task and not

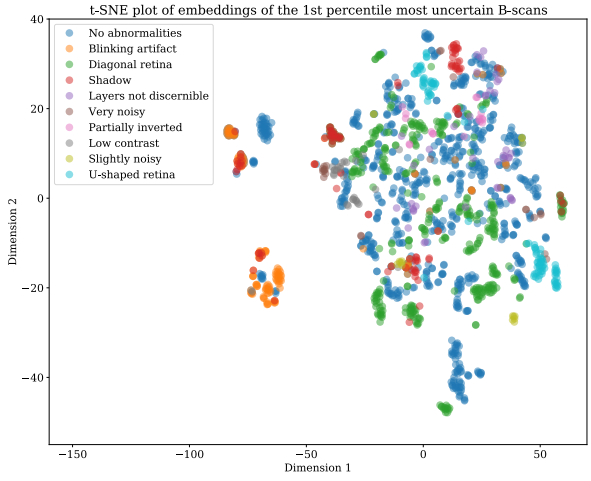


Figure 5.A.1: t-SNE plot of the embeddings of the first percentile of B-scans with the highest ordinal entropy (n = 1,516). The types of artifacts and abnormalities related to image acquisition are colored differently.

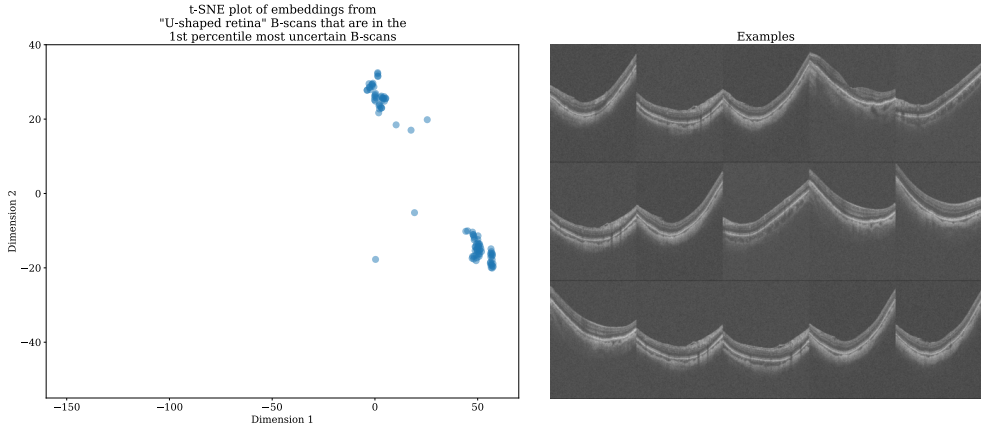


Figure 5.A.2: t-SNE plot and examples for "U-shaped retina" B-scans. There were 99 "U-shaped retina" B-scans from 8 OCT volumes.

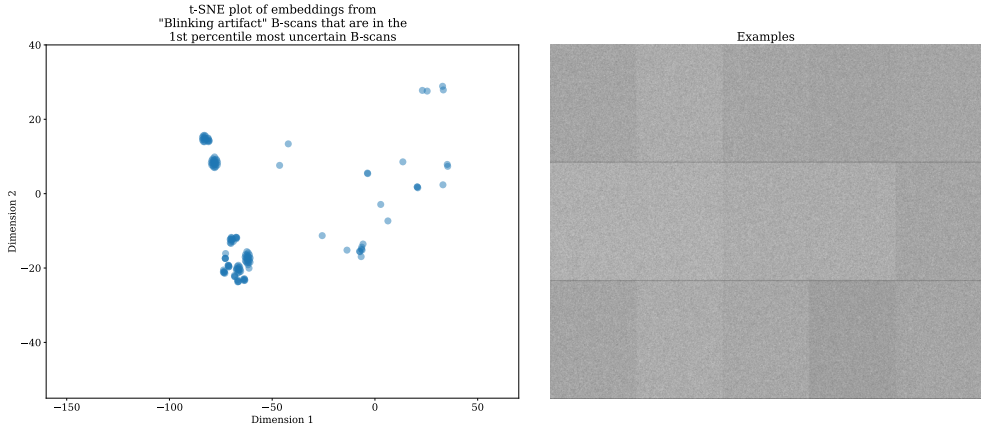


Figure 5.A.3: t-SNE plot and examples for "Blinking artifact" B-scans. There were 108 "Blinking artifact" B-scans from 14 OCT volumes.

for artifact classification. However, some artifact types are relatively isolated from others, such as "Blinking artifacts" and "U-shaped retina".

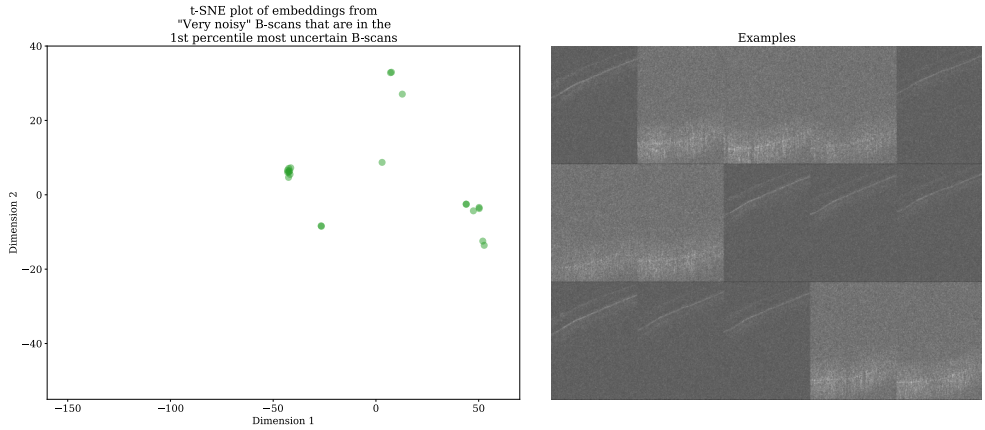


Figure 5.A.4: t-SNE plot and examples for "Very noisy" B-scans. There were 22 "Very noisy" B-scans from 2 OCT volumes.

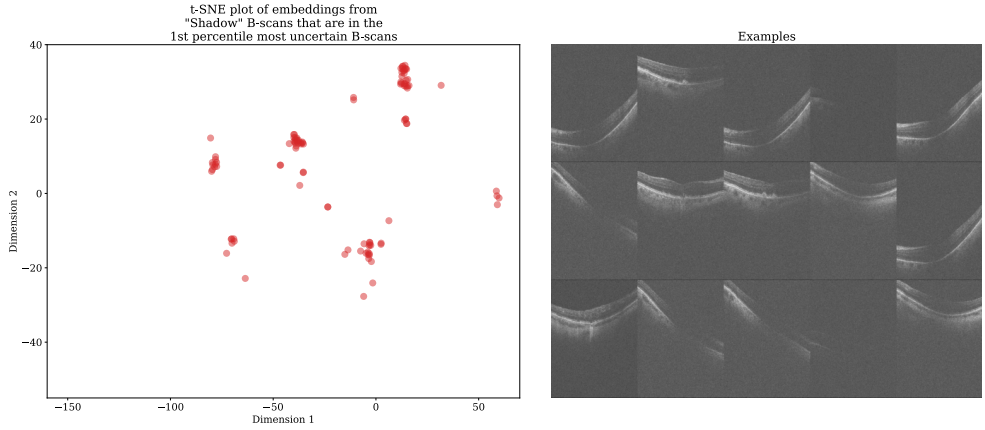


Figure 5.A.5: t-SNE plot and examples for "Shadow" B-scans. There were 82 "Shadow" B-scans from 16 OCT volumes.

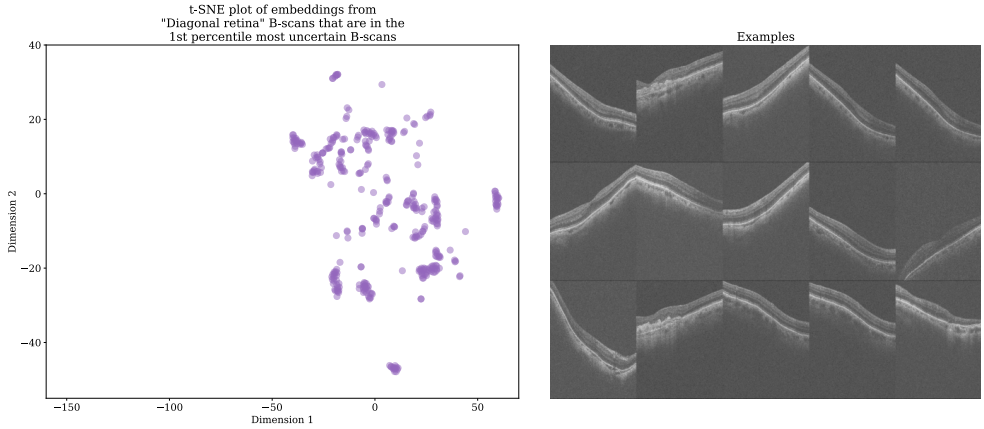


Figure 5.A.6: t-SNE plot and examples for ”Diagonal retina” B-scans. There were 357 ”Diagonal retina” B-scans from 21 OCT volumes.

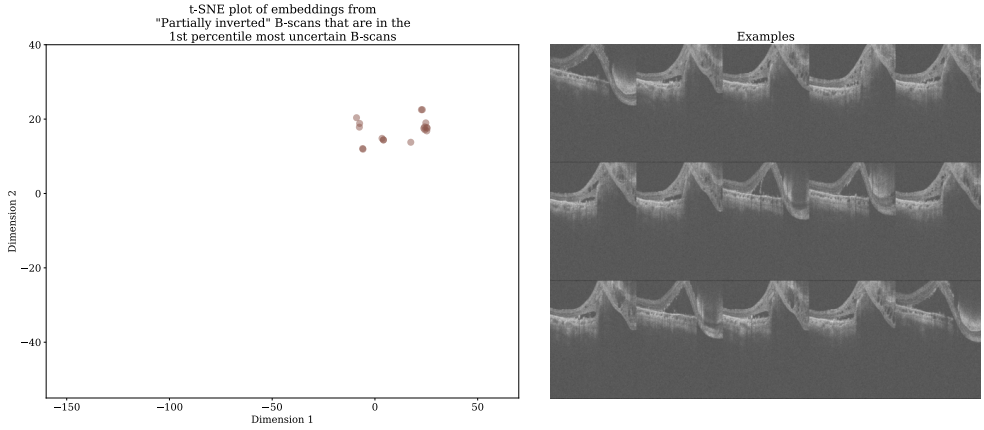


Figure 5.A.7: t-SNE plot and examples for ”Partially inverted” B-scans. There were 18 ”Partially inverted” B-scans from 1 OCT volumes.

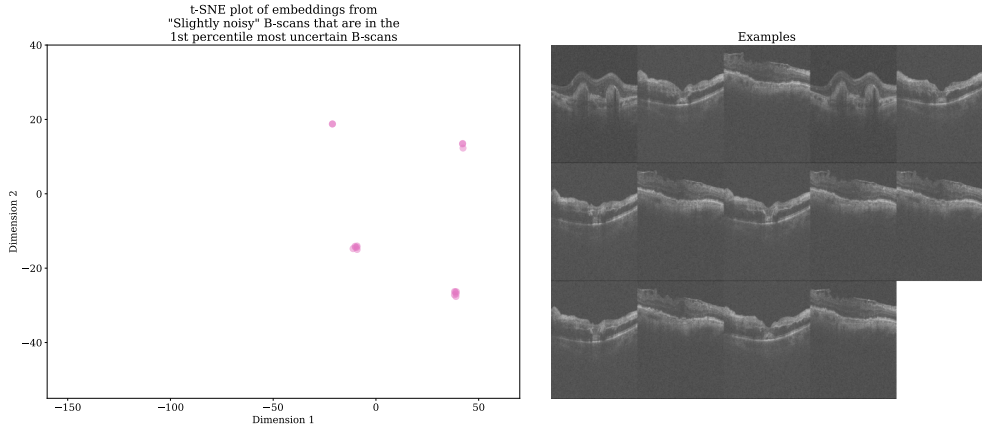


Figure 5.A.8: t-SNE plot and examples for ”Slightly noisy” B-scans. There were 14 ”Slightly noisy” B-scans from 3 OCT volumes.

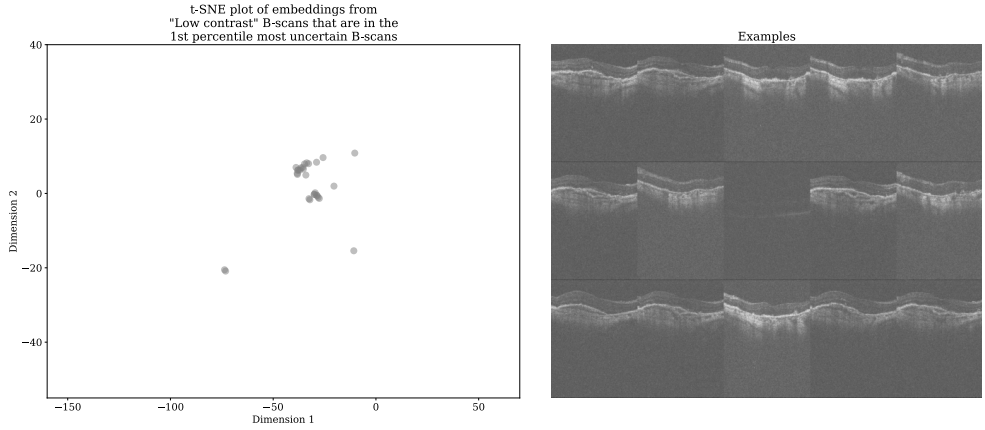


Figure 5.A.9: t-SNE plot and examples for ”Low contrast” B-scans. There were 29 ”Low contrast” B-scans from 3 OCT volumes.

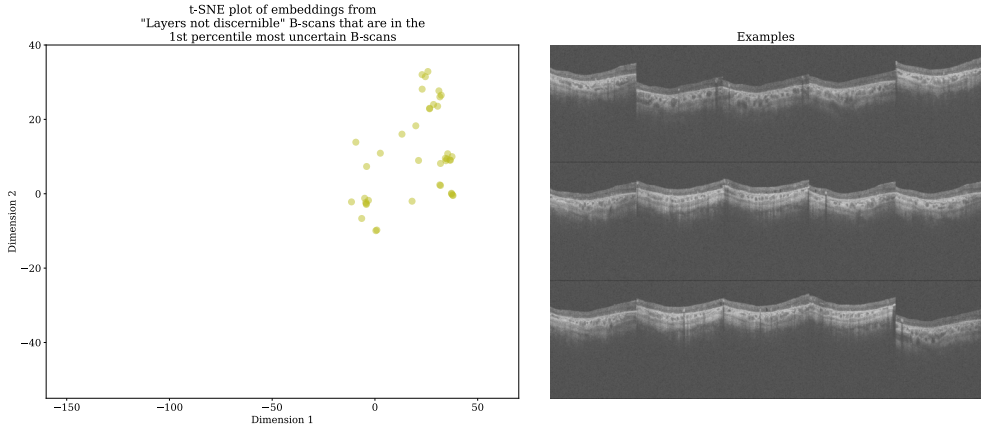


Figure 5.A.10: t-SNE plot and examples for "Layers not discernible" B-scans. There were 41 "Layers not discernible" B-scans from 1 OCT volumes.

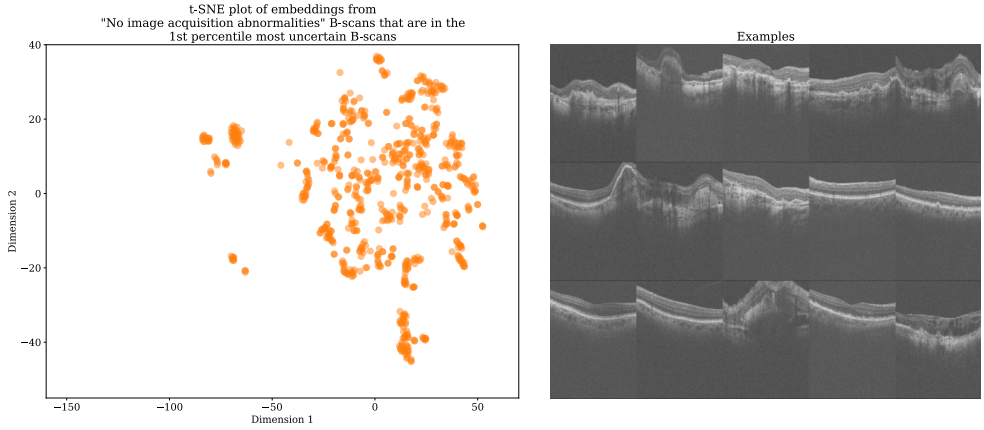


Figure 5.A.11: t-SNE plot and examples for "No image acquisition abnormalities" B-scans. There were 746 "No image acquisition abnormalities" B-scans from 55 OCT volumes.

6

Generalizable iRORA and cRORA detection models

Authors: Coen de Vente, Philippe Valmaggia, Carel B. Hoyng, Frank G. Holz, Mohammad M. Islam, Caroline C. W. Klaver, Camiel J. F. Boon, Steffen Schmitz-Valckenberg, Adnan Tufail, Marlene Saßmannshausen*, Clara I. Sánchez* on behalf of the MACUSTAR Consortium† (*equal contribution) (†see chapter acknowledgements)

Original title: Generalizable Deep Learning for the Detection of Incomplete and Complete Retinal Pigment Epithelium and Outer Retinal Atrophy: A MACUSTAR Report

Published in: Translational Vision Science & Technology (Volume: 13, Issue: 9, Pages: 11, September 2024)

https://doi.org/10.1167/tvst.13.9.11

Abstract

Purpose: To develop a deep learning algorithm for detecting and quantifying incomplete (i) and complete (c) retinal pigment epithelium and outer retinal atrophy (RORA) in optical coherence tomography (OCT) that generalizes well to data from different devices and to validate in an intermediate age-related macular degeneration (iAMD) cohort.

Methods: The algorithm comprised a domain adaptation (DA) model, promoting generalization across devices, and a segmentation model for detecting granular biomarkers defining iRORA/cRORA, which are combined into iRORA/cRORA segmentations. Manual annotations of iRORA/cRORA in OCTs from different devices in the MACUSTAR study (168 iAMD patients) were compared to the algorithm's output. Eye level classification metrics included sensitivity, specificity, and quadratic weighted Cohen's κ score (κ_w). Segmentation performance was assessed quantitatively using Bland-Altman plots and qualitatively.

Results: For ZEISS OCTs, sensitivity and specificity for iRORA/cRORA classification were 38.5% and 93.1%, respectively, and 60.0% and 96.4% for cRORA. For Spectralis OCTs, these were 84.0% and 93.7% for iRORA/cRORA, and 62.5% and 97.4% for cRORA. κ_w scores for three-way classification (None, iRORA, cRORA) were 0.37 and 0.73 for ZEISS and Spectralis, respectively. Removing DA reduced κ_w from 0.73 to 0.63 for Spectralis.

Conclusions: The DA-enabled iRORA/cRORA segmentation algorithm showed superior consistency compared to human annotations, and good generalization across OCT devices.

Translational Relevance: The application of this algorithm may help towards precise and automated tracking of iAMD-related lesion changes, which is crucial in clinical settings and multicenter longitudinal studies on iAMD.

6.1 Introduction

The robust and reliable detection of structural biomarkers for early disease progression in age-related macular degeneration (AMD) is crucial for validating potential structural endpoints in future interventional trials [266–268]. Alongside the recent emergence of fast, highly reproducible, and high-resolution multimodal retinal imaging, various intermediate AMD (iAMD) biomarkers for early atrophy manifestation have been identified based on optical coherence tomography (OCT) [269–274]. This advancement has enabled more accurate disease progression predictions in individual patients [275–278]. The Classification of Atrophy Meetings (CAM) group, comprised of international AMD and retinal imaging specialists, has defined OCT imaging-based lesions of atrophy development as incomplete retinal pigment epithelium and outer retinal atrophy (iRORA) and complete retinal pigment epithelium and outer retinal atrophy (cRORA) [62,279]. Ongoing and potentially upcoming observational iAMD studies focus on utilizing these lesions of atrophy development [266,280,281].

Despite these advancements, the improvement of multimodal retinal imaging data presents practical challenges, as human assessment alone cannot keep up with the large multicenter datasets that are becoming available. Therefore, we need automated approaches to support image grading and data analysis. Various studies have presented deep learning models for detecting iRORA and cRORA in OCT scans [68,282–284]. These algorithms were all evaluated using OCT data from a single device (Heidelberg Spectralis), which was also used to acquire their training data.

However, many studies investigating AMD progression and real-world clinical settings use a wide variety of OCT devices [56]. At the same time, deep learning models trained on data from one OCT device have been shown to generalize poorly to other devices without any further training [205, 242, 285], illustrating the wider challenge of compromised deep learning reliability under data shift in computer vision [208] and medical imaging [286–289]. Automated approaches to localize iRORA, cRORA, and their underlying granular features that generalize well to data from different device manufacturers are still warranted. A traditional fully-supervised learning approach would require extensive manual annotation for the data of each OCT device, a process that is both time-intensive and resource-demanding.

This study aims to develop and evaluate a machine learning pipeline for detecting the underlying structural characteristics defining iRORA and cRORA lesions in OCT that is robust to data from two different devices, without requiring manually annotated training data from those devices. We use the MACUSTAR study, a prospective, multicenter, and

low-interventional clinical study in subjects with AMD, for evaluation [266]. It contains standardized OCT data acquired with two common devices from different manufacturers, enabling the rigorous assessment of model generalizability across these devices.

6.2 Methods

6.2.1 Dataset and study cohort

Evaluation data for iRORA and cRORA detection algorithm: The MACUSTAR study cohort

MACUSTAR is a prospective, multicenter, and low-interventional clinical study (ClinicalTrials.gov Identifier: NCT03349801) in subjects with AMD that is conducted at 20 sites in 7 European countries and consists of a cross-sectional and longitudinal study part. Detailed information on the design and clinical study protocol of the study have been reported previously [266]. Human research ethics committee approval was obtained at all participating clinical sites before enrollment, complying with all applicable legal regulations as previously described. Participants provided informed consent before study recruitment and data collection, and this study has been conducted according to the provisions of the Declaration of Helsinki.

The study's inclusion and exclusion criteria have been described previously in detail [290,291]. In brief, based on Ferris *et al.*, iAMD was defined by the presence of large drusen (> 125 µm) and/or presence of pigmentary abnormalities associated with medium drusen (> 63 µm and ≤ 125 µm) [61]. Extrafoveal atrophy (< 1.25 mm^2) outside the central subfield of the Early Treatment Diabetic Retinopathy Study Grid was permitted in the fellow eye of the iAMD study eye. If there was any geographic atrophy (> 0.1 mm^2 in fundus autofluorescence imaging) or macular neovascularization the study patient was included in the late AMD group of MACUSTAR. The herein presented analysis is only based on iAMD study eyes of the cross-sectional study part of the baseline visit (V2) of MACUSTAR.

6.2.2 Retinal imaging protocol of the MACUSTAR study

Following pupil dilation (tropicamide 0.5% and phenylephrine 2.5%) MACUSTAR study patients underwent multimodal retinal imaging according to standard operational procedures by certified study site personnel. The retinal imaging protocol comprised spectral domain (SD) OCT imaging performed with the Cirrus (Macular Cube 200 × 200, centered on fovea, signal strength at least 6/10, Zeiss, Meditec, Oberkochen, Germany) and the

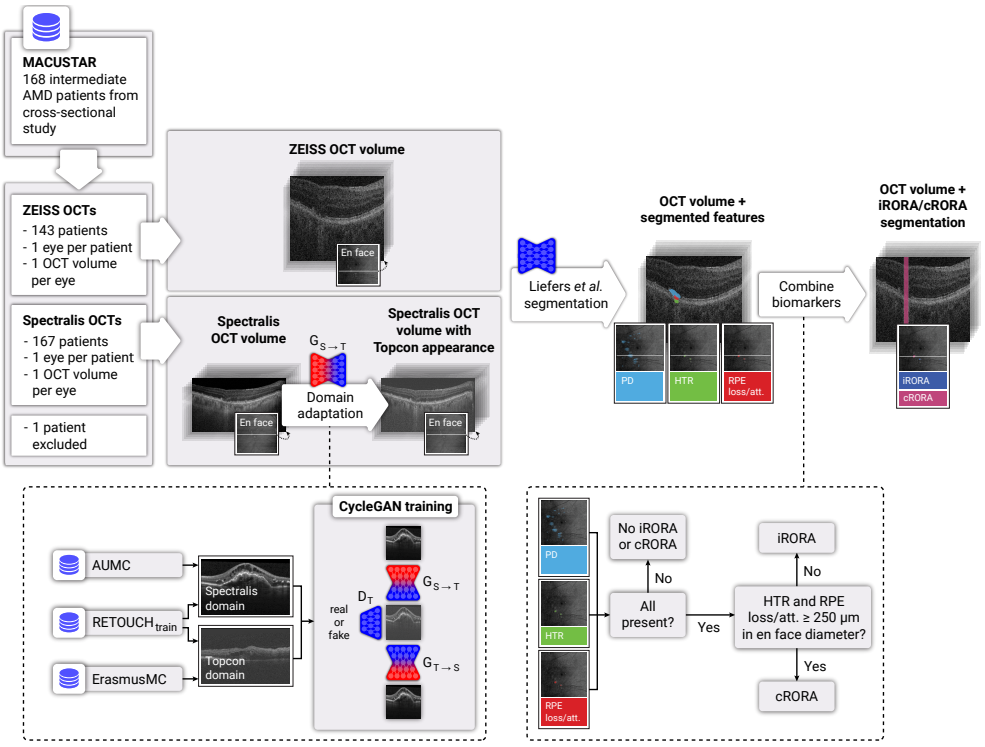


Figure 6.1: Overview of the methods used for the automated iRORA/cRORA detection pipeline. Spectralis data was fed into a domain adaptation model. The transformed OCTs were subsequently used as input for the Liefers *et al.* [68] segmentation model. ZEISS OCTs were directly fed into the segmentation model. The OCT volumes were fully segmented for all segmented features, including iRORA/cRORA. In the CycleGAN training diagram, $G_{S\rightarrow T}$ is the generator for transforming the appearance of Spectralis images to the appearance of Topcon OCTs. $G_{T\rightarrow S}$ is the generator for transforming the appearance of Topcon images to the appearance of Spectralis OCTs. DT is the discriminator for the Topcon image domain. HTR: hypertransmission. att.: attenuation. PD: photoreceptor degeneration.

Spectralis device (HRA + OCT device, digital imaging resolution 768 × 768 pixels, 30° × 25°, enhanced-depth-imaging mode, centered on the fovea, 241 B scans, distance between scans 30 µm, automatic real-time mode, 9 frames, Heidelberg Engineering, Heidelberg, Germany). For each patient, the Spectralis and ZEISS OCTs were acquired at the same study date and during the same visit. Further details on the standardized retinal imaging protocol with additionally acquired imaging modalities within the MACUSTAR study have been previously reported [290, 291]. All imaging data were transmitted to the GRADE Reading Center (University of Bonn, Germany).

6.2.3 Annotation of iRORA and cRORA lesions

Manual annotation of iRORA and cRORA lesions was performed at the B-scan level in both the Zeiss and Spectralis SD-OCT imaging datasets by readers experienced in AMD studies at the GRADE Reading Centre, who had received detailed training in iRORA/cRORA detection. Therefore, each OCT B-scan of both imaging devices was reviewed and carefully screened for meeting predefined lesion criteria. According to the definition by the CAM group [62], a lesion was graded and annotated as iRORA if there was: (1) a region of choroidal hypertransmission < 250 µm in diameter and (2) a zone of attenuation or disruption of the retinal pigment epithelium (RPE) of < 250 µm in diameter, (3) evidence of photoreceptor degeneration, and (4) absence of an RPE tear. A lesion was defined as cRORA, if the aforementioned size criteria related to choroidal hypertransmission and RPE attenuation or disruption exceeded 250 µm [279]. This diameter was measured automatically by calculating the Feret diameter (in any direction in the *en face* plane, using scikit-learn [292]). Manual annotation of iRORA/cRORA was performed in all B-scans meeting lesion criteria. The screening of OCT B-scans as well as the annotation of iRORA/cRORA lesions was first performed in the 200 × 200 Macular Cube SD-OCT scans of each study patient, followed by the 241 B-scans Spectralis SD-OCT imaging dataset. The grading and annotation of iRORA/cRORA lesions were performed on the Zeiss SD-OCT imaging data in the device's manufacturer platform (Zeiss Forum, Retina Workplace). For the manual annotation of iRORA/cRORA lesions in the Spectralis data set, the 241 B-scan SD-OCT imaging data were imported for each study patient in a raw file format to an at the reading center developed annotation platform based on the Jupyter Notebook platform (*jupyter.org*). Under detailed consideration of each OCT B-scan, OCT volumes of both manufactures were graded for the presence of iRORA/cRORA by a medical reader who is highly experienced in AMD trials. If a lesion was detected as present, the maximum horizontal diameter of the lesion was manually annotated in each B-scan fulfilling lesion criteria using the highlighting tools implemented in the corresponding platforms. Annotated imaging data were subsequently exported for further detailed AI-based data analysis.

In a sub-analysis, manually performed annotations were revised in 12 study eyes with SD-OCT imaging based on discrepancies between the artificial intelligence (AI) model outputs and manual references.

Development data for domain adaptation For the development of the DA algorithm, we collected macula-centered SD-OCTs that were independent of the MACUSTAR data for both the Spectralis and Topcon domains. An overview of these datasets is shown in Table 6.1.

The data from the Spectralis domain originated from the RETOUCH dataset [66] and

Table 6.1: Overview of the development data of the domain adaptation model.

		Training			Validation		
Domain	Source	Patients	OCT volumes	B-scans	Patients	OCT volumes	B-scans
Spectralis	AUMC	20	20	446			
	RETOUCH	24	24	1,176	14	14	686
Topcon	RETOUCH	21	21	2,688	13	13	1,664
	Rotterdam Study	100	97	12,800			

a proprietary dataset from the Amsterdam University Medical Centers (AUMC). RETOUCH contains data from patients with AMD, retinal vein occlusion, and diabetic macular edema. From the RETOUCH dataset, 24 Spectralis OCT volumes were used for the training data and 14 for the validation data. The RETOUCH OCTs contained 49 B-scans per volume with a size of 512 × 496 or 1024 × 496 pixels, in a raster of 6 × 6 mm^2. From the AUMC dataset, 10 patients with AMD and 10 patients with X-linked retinoschisis were included in the training set. The AUMC OCTs contained either 19 or 25 B-scans per volume with a size of either 512 × 496 or 1024 × 496 pixels, in a raster of approximately 6 × 6 mm^2. For the Spectralis domain, this resulted in a total of 1622 B-scans from 44 patients for training and 686 B-scans from 14 patients for validation.

The Topcon domain data originated from the RETOUCH [66] dataset and the Rotterdam Study [251]. The Rotterdam Study is a prospective cohort study established in 1990 in the city of Rotterdam, the Netherlands. From this study, 100 Topcon OCT volumes were randomly selected for the training set of the Topcon domain. This resulted in a set of 100 OCTs from 97 patients, as we included two OCTs from three patients and one OCT from all other patients. From the RETOUCH training set, 21 Topcon OCT volumes were used for the training set of the Topcon domain and 13 for the validation data. The Topcon domain OCTs were acquired using T-1000 or T-2000 devices, each containing 128 B-scans of 512 × 885 or 512 × 650 pixels in a raster of 6 × 6 mm^2. For the Topcon domain, this resulted in a total of 15488 B-scans from 118 patients for training and 1664 B-scans from 13 patients for validation.

6.2.4 iRORA and cRORA detection model

For the detection of iRORA and cRORA in OCT scans, we processed each OCT volume with our AMD biomarker segmentation model [68], obtaining the volumetric segmentation and quantification of 13 different structural biomarkers. The biomarker segmentation model was a deep learning model based on U-Net [9]. This model took 9 consecutive B-scans as input, providing segmentations for the middle B-scan. It was a convolutional neural network with 3 × 3 × 1 convolutions in U-Net layers and 1 × 1 × 3 convolutions in the deeper

layers of the encoder. Following the iRORA/cRORA definitions defined by the CAM group [62, 279], we combined three of these structural features, namely hypertransmission, RPE loss or attenuation, and ellipsoid zone loss (as a criterion for photoreceptor degeneration), as shown in Fig. 6.1. If the three features detected by the biomarker segmentation model co-located in the A-scan (vertical) direction, the A-scans of that lesion were classified as either iRORA or cRORA. If these three features did not co-locate, the A-scan was classified as background. Subsequently, if the hypertransmission and RPE loss of this potential iRORA/cRORA lesion had a Feret diameter (in the *en face* plane, calculated using scikit-learn [292]) of at least 250 µm, the lesion was considered a cRORA lesion. If the diameter was less than 250 µm, it was considered a lesion of iRORA.

Pre-processing of the Spectralis data was done using a domain adaptation (DA) approach, as described in the next section. For compatibility with the segmentation model, all 3D OCT volumes were cropped or padded to $6 \times 2.3 \times 6$ mm^3 around the volume center. Subsequently, we resized the volumes to $128 \times 650 \times 512$ voxels, preparing them for input into the segmentation model.

6.2.5 Domain adaptation

The Spectralis OCT volumes were transformed using DA, based on Cycle-Consistent Generative Adversarial Networks (CycleGAN) [293]. We aimed to modify the appearance of the Spectralis scans to align with the appearance of Topcon OCT scans, intending to promote compatibility with the training data distribution of the biomarker segmentation model (trained on Topcon images). We did not employ DA in the ZEISS OCT processing pipeline, as these OCTs already showed very similar image characteristics to the Topcon OCTs used to train the biomarker segmentation model. This is supported by the settings used in the ZEISS and Topcon devices, which do not use B-scan averaging, a scanning protocol that substantially impacts the signal-to-noise ratio. In contrast, this protocol is used in the Heidelberg Spectralis device [214].

CycleGANs are designed for image-to-image translation, allowing the translation of an image from a source domain to a target domain without paired data. It uses two generators, one for generating images from the Spectralis domain to the Topcon domain and one for generating images from the Topcon domain to the Spectralis domain. Moreover, one discriminator for each domain is trained to distinguish real images from fake images. The aim of the generators is to fool the discriminators into classifying real images as fake. Therefore, for validation and application of the CycleGAN in the iRORA/cRORA detection pipeline, we only used the generator for transforming Spectralis into Topcon images. As a pre-processing step, all Spectralis OCT volumes were resized to $512 \times 650 \times 128$ pixels to

ensure compatibility with our biomarker segmentation model.

We largely employed the same network and optimization settings as described in the original CycleGAN paper. We used the Adam optimizer [153] with a learning rate of 2×10^{-4}. The model was trained for 85 epochs (1,316,480 iterations). The adversarial loss was implemented using the mean squared error. The cycle consistency loss and the identity loss were implemented using the mean absolute error. The CycleGAN was trained on 2D patches of 256 × 256 pixels, randomly selected from the pre-processed OCTs. When generating images during test time, the whole B-scans of 512 × 650 pixels were used as model input.

6.2.6 Statistical analysis

We evaluated the agreement of iRORA/cRORA detection between the deep learning model and manual annotations on both eye level and in terms of area. To convert voxel-level annotations and outputs of iRORA/cRORA to eye level, we employed the following workflow. If any cRORA was found, the eye level class became cRORA. If no cRORA was present, but iRORA was found, the eye level class became iRORA. Otherwise, the eye level label was None.

For the eye level analysis, we evaluated the agreement using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and quadratic weighted Cohen's κ score (κ_w). The sensitivity and specificity were calculated both with an onset of iRORA (None vs. iRORA/cRORA) and with an onset of cRORA (None/iRORA vs. cRORA). κ_w was calculated on the three-category classification problem of None vs. iRORA vs. cRORA. We reported the agreement between the human grader and the AI model using the aforementioned metrics. Furthermore, we assessed the intra-rater agreement for the human grader and the AI model separately using κ_w. When calculating standard deviations and p-values for κ_w, we used non-parametric bootstrapping [152] with 1,000 iterations. We calculated the agreement in terms of area using Bland-Altman plots.

The domain adaptation model was both qualitatively and quantitatively validated. For the quantitative analysis, we computed the Dice score of the fluid segmentation performance of a 2D nnU-Net [27] model that was trained on the Topcon data from RETOUCH. The Dice score was calculated separately for each OCT volume. The mean and standard deviation of these Dice scores were subsequently calculated. The qualitative analysis was performed by visually inspecting artificially generated Topcon examples from real Spectralis images.

6.3 Results

The cross-sectional part of the MACUSTAR study included a total of 168 study eyes of 168 patients with iAMD (mean age ± standard deviation: 71.2 ± 7.55 years). Due to the absence of 25 ZEISS 200 × 200 Macular Cube SD-OCT scans and insufficient image quality of one 241 B-scan Spectralis SD-OCT volume at baseline visit (V2), a total number of 143 iAMD patients with ZEISS Macular Cube SD-OCT, and 167 iAMD patients with Spectralis SD-OCT imaging were included for further analysis.

6.3.1 Segmentation and eye level classification performance

Table 6.2 presents the performance metrics of the iRORA/cRORA detection pipeline for eye-level classification, evaluated using the ZEISS and Spectralis datasets. The table includes comparisons both with and without the application of the DA approach to the Spectralis data. The specificities were high for both OCT devices and onsets, ranging from 93.1% to 97.4%. The sensitivities were lower, ranging from 38.5% to 84.0%. The κ_w between the manual annotations and the AI model was substantially higher for the Spectralis OCTs (0.73) than for the ZEISS OCTs (0.37). Fig. 6.2 shows the confusion matrices for the eye level evaluation. Qualitative segmentation results are shown in Figures 6.3 and 6.4 for the Spectralis and ZEISS data, respectively. The figures illustrate scenarios of both high agreement between the model and the manual annotations, but also different types of disagreement, in the form of a false negative and a false positive.

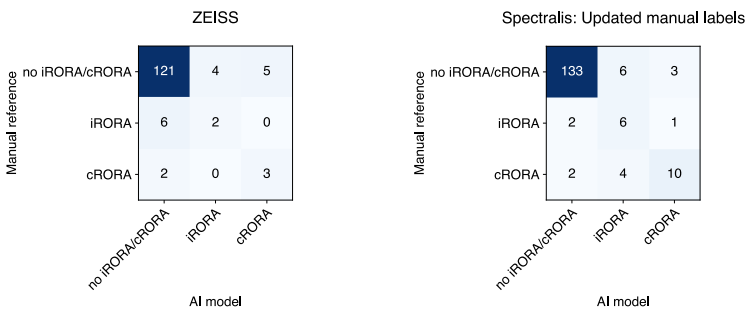
For most cases, OCT volumes from the ZEISS and Spectralis devices were available. The AI model and human grader interpreted these independently, blinded to the OCT from the other device. The discrepancies between the corresponding eye level outputs when interpreting the volumes from these two devices are visualized in Appendix Fig. 6.A.1. In Appendix Fig. 6.A.2, Bland-Altman plots to compare the predicted areas of the manual and AI segmentations of the iRORA and cRORA lesions are shown separately. An analysis of the manual annotations that were updated based on the discrepancies between the initial version of the annotations and the model predictions can be found in Appendix Fig. 6.A.3. An example of such an update is shown in Appendix Fig. 6.A.4.

Effect of domain adaptation

The effect on κ_w (between the manual annotations and the AI model) and the confusion matrix of using DA in the iRORA/cRORA detection pipeline for the Spectralis data is shown in Fig. 6.7. The κ_w score was 0.73 with DA, while it was 0.63 without DA. This difference

Table 6.2: Performance for eye level iRORA and cRORA classification. n is the number of OCT volumes, which is equal to the number of eyes and the number of patients. PPV: positive predictive value. NPV: negative predictive value.

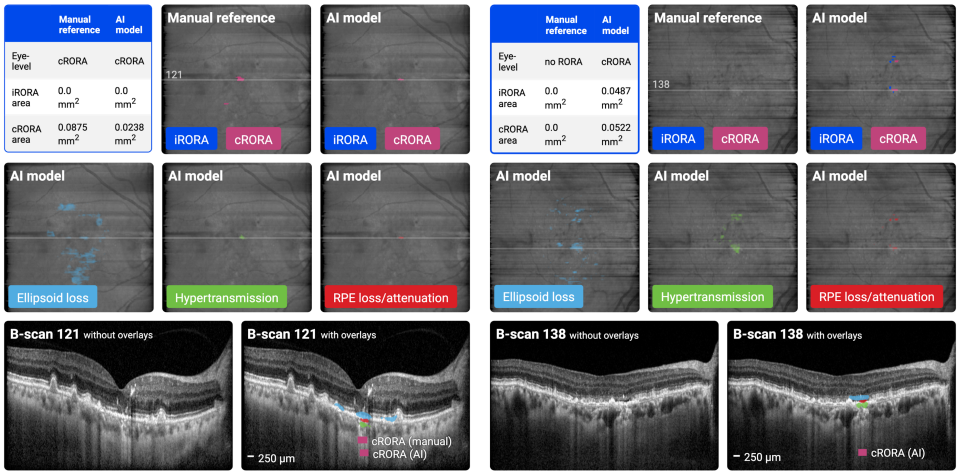
		OCT device	
		ZEISS (n = 143)	Spectralis (n = 167)
None vs. iRORA/cRORA	Sensitivity	38.5%	84.0%
	Specificity	93.1%	93.7%
	PPV	35.7%	70.0%
	NPV	93.8%	97.1%
None/iRORA vs. cRORA	Sensitivity	60.0%	62.5%
	Specificity	96.4%	97.4%
	PPV	37.5%	71.4%
	NPV	98.5%	96.1%
None vs. iRORA vs. cRORA	κ_w	0.37	0.73



(a) (b)

Figure 6.2: Confusion matrices, comparing the manual reference with the AI model for the images originating from the **(a)** ZEISS and the **(b)** Spectralis datasets.

was not statistically significant (p = 0.087). There were 6 cases with cRORA according to the manual reference which were classified as no iRORA/cRORA by the model without DA. This was reduced to 2 cases when including DA in the approach. However, the number of cases that the model classified as cRORA but were negative according to the manual reference did increase from 1 to 3 when introducing DA into the pipeline. An example where the model with DA picked up an iRORA lesion that the model without DA missed, is shown in Fig. 6.8. The validation results of the DA model when used in a fluid segmentation pipeline on the RETOUCH test set are shown in Table 6.A.1 of the Appendix.



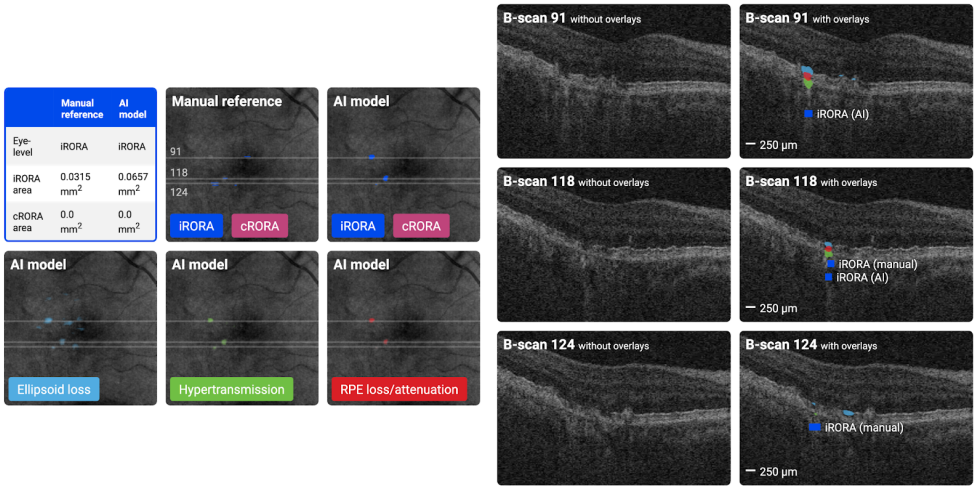
(a) (b)

Figure 6.3: Examples of Heidelberg Spectralis results. Each subfigure shows a table in the top left, containing the eye-level decisions and the *en face* iRORA/cRORA areas, calculated from the segmentations, for both the manual reference and the AI model. The middle and right cells of the first rows show the iRORA/cRORA segmentations from the manual reference and the AI model, respectively, overlaid on the *en face* projection of the OCT volume. The middle rows show the segmentations from the AI model for the three underlying granular biomarkers of iRORA/cRORA, overlaid on the *en face* projection of the OCT volume. Each *en face* image shows a white horizontal line, which corresponds to the B-scan shown in the last row. The last rows show the AI model segmentation of the granular biomarkers. If iRORA/cRORA is detected, the last rows also show horizontal lines labeled as either "iRORA/cRORA (AI)" for the AI model or "iRORA/cRORA (manual)" for the manual reference. **(a)** True positive for cRORA. **(b)** False positive example for cRORA. These cases exhibit a mixed phenotype with presence of sub-RPE drusen and subretinal drusenoid deposits as well as evidence of an overall on the B-scan increased choroidal hypertransmission, leading to ambiguity for the human reader when assessing the presence of iRORA/cRORA.

6.4 Discussion

In this study, we presented a generalizable deep learning pipeline for the detection of iRORA/cRORA lesions in OCT. By using semantic segmentation of the underlying granular biomarkers, rather than direct detection of these lesions, we aimed to be resilient to possible definition changes of iRORA/cRORA. We employed DA in the pipeline to ensure applicability to other devices beyond those used during training. We assessed the performance of this approach in iAMD eyes of the cross-sectional part of the MACUSTAR study, a prospective, multicenter, and low-interventional clinical study with data from multiple OCT devices per visit and patient, by comparing the model outputs to human annotations in a reading center setting.

For both OCT devices, and the eye level classification with an onset of both iRORA and



	Manual reference	AI model
Eye-level	iRORA	iRORA
iRORA area	0.0315 mm^2	0.0657 mm^2
cRORA area	0.0 mm^2	0.0 mm^2

Figure 6.4: An example of ZEISS results. A table is shown in the top left, containing the eye-level decisions and the *en face* iRORA/cRORA areas, calculated from the segmentations, for both the manual reference and the AI model. The middle and right cells of the first row on the left show the iRORA/cRORA segmentations from the manual reference and the AI model, respectively, overlaid on the *en face* projection of the OCT volume. The middle row on the left shows the segmentations from the AI model for the three underlying granular biomarkers of iRORA/cRORA, overlaid on the *en face* projection of the OCT volume. Each *en face* image shows white horizontal lines, which correspond to the B-scans on the right. On these B-scans, the AI model segmentation of the granular biomarkers. If iRORA/cRORA is detected, these B-scans also show horizontal lines labeled as either "iRORA/cRORA (AI)" for the AI model or "iRORA/cRORA (manual)" for the manual reference. B-scan 91 contains a false positive, where all three features were segmented by the AI model. In B-scan 118, the model and the AI model both detected the same iRORA lesion. B-scan 124 contains an iRORA lesion according to the manual reference. The model did find very small areas of ellipsoid loss and hypertransmission at this location, but RPE loss/attention was not segmented. Therefore, iRORA was not identified by the model.

cRORA, the specificity of the AI model was substantially higher than the sensitivity. The Bland-Altman plots indicate almost no systematic bias, but show a number of outliers that are mostly due to the AI model segmenting larger areas than the human grader. Fig. 6.4 illustrates this phenomenon, where both the human grader and the AI model found iRORA lesions, but the total *en face* area computed from the AI model segmentations was more than twice as large as the area computed from the manual annotations.

Unambiguous manual delineation of lesions appeared particularly challenging in eyes with mixed drusen phenotypes of subretinal drusenoid deposits and sub-RPE drusen as well as evidence of an overall increased choroidal hypertransmission on the OCT line scan (cp. Fig. 6.3b). Additional *en face* assessment of choroidal hypertransmission can further help to delineate these lesions.

The AI model performance was generally higher on the Spectralis data than on the ZEISS data. As can be observed in Table 6.2, the κ_w score was 0.73 for the Spectralis data,

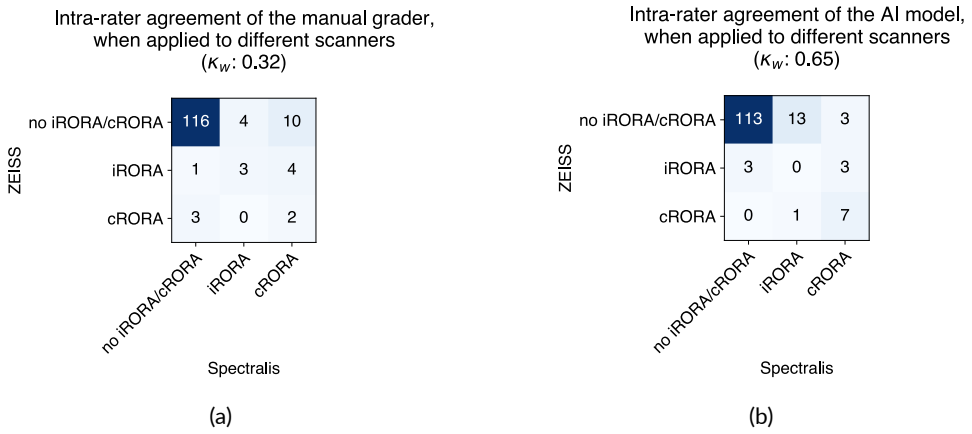


Figure 6.5: Confusion matrices showing the intra-rater agreements of both **(a)** the manual grader and **(b)** the AI model when comparing their linked decisions on OCT-volume level for the ZEISS and the Spectralis data.

while it was 0.37 for the ZEISS data. This might be a consequence of the lower signal-to-noise ratio of the latter OCT device, causing the task of reliably detecting iRORA/cRORA lesions to be more challenging for both AI models and human graders. This can lead to a reference standard that is further away from the ground truth, causing performance metrics to be lower. Furthermore, the Spectralis results depicted in Table 6.2 were based on the manual labels that were updated based on the discrepancies between human and manual labels. However, the κ_w score computed when using the original manual labels was 0.53, which was still higher than for the ZEISS OCTs. The use of DA for the Spectralis data, which was not employed to the ZEISS data, could also explain these differences, despite the high similarity between the ZEISS and training data (as described in the subsection "Domain Adaptation" of the "Methods" section). In future work, our assumption that DA was not necessary for the ZEISS data could be tested empirically.

In the Spectralis dataset, the cases with a discrepancy between the AI model and initial manual labels on eye level were reconsidered manually. This sub-analysis resulted in an updated label for 12 cases out of a total of 25 cases with iRORA/cRORA in the final manual labels. 8 out of these 12 cases were positive iRORA/cRORA cases that were initially a less severe classification. This indicates the necessity of AI-based workflows in future clinical settings and trials with iRORA/cRORA endpoints. Using AI to highlight missed cases or presenting AI findings before a human starts grading can potentially increase clinical accuracy. Therefore, we think initiating a discussion in the research and clinical community on how to incorporate AI systems in these settings is highly important.

We evaluated the consistency of both the deep learning model and the human grader

across different OCT devices. This intra-rater agreement was substantial for the AI model (κ_w = 0.65) and fair for the manual grader (κ_w = 0.32). Nevertheless, there still appears to be room for improvement in both cases, indicating the need for standardized OCT data. DA approaches, such as the one described in this work, may contribute towards this goal of data standardization.

An increase in terms of κ_w from 0.63 to 0.73 was shown when introducing DA into the processing pipeline of the Spectralis data. However, this difference was not statistically significant (p = 0.087), which may be caused by the relatively small number of iRORA/cRORA cases in the evaluation set (n = 23 according to the manual labels, and n = 30 according to the deep learning model).

Several previous studies have described the development and evaluation of algorithms for iRORA/cRORA classification or segmentation [282–284]. A direct performance comparison of our deep learning framework to these other algorithms is challenging due to the lack of benchmark datasets and standardized evaluation protocols. For example, Chiang *et al.* [282] described a deep learning approach for iRORA/cRORA classification of individual B-scans. Their approach achieved higher sensitivities than ours, but our method demonstrated higher performance in terms of specificity, PPV, and NPV. However, a direct comparison of these classification metrics is problematic for several reasons. (1) The operating points used in their and our work lead to highly different trade-offs between sensitivity, specificity, PPV, and NPV. Therefore, it would be valuable to perform a comparison between these algorithms using receiver operating characteristic (ROC) curves in future work. We did not present ROC curves in this paper because our current algorithm does not provide probabilities for the detected iRORA/cRORA lesions, as this detection is based on the binary presence of the underlying features. (2) We reported the iRORA/cRORA (*i.e.,* the two classes of iRORA and cRORA merged) and cRORA classification performance, while they reported the classification performance of iRORA and cRORA separately. (3) For some of their used datasets, they only reported metrics for classification on B-scan level instead of eye level. (4) There is a large difference in the type of populations used between the datasets used in their and our work, leading to highly varying class imbalances and distributions of AMD stages.

Another work that presented a deep learning model for iRORA/cRORA detection is described by Szeskin *et al.* [283] They only reported metrics for column classification, further hampering a one-to-one comparison with the results of this study. Zhang *et al.* [284] presented an automatic quantification approach for iRORA/cRORA as well, but they also did not provide metrics on eye level for iRORA/cRORA classification.

An important strength of this study was the large evaluation dataset, which included

high-quality imaging OCT data and was prospectively collected as part of a multicenter study. The OCT data from both acquisition devices used a highly dense OCT raster. This ensured that small lesions were captured in each scan, which is important in the quantification of lesions related to iAMD, such as iRORA. Furthermore, the human annotator was highly experienced with reading retinal images within a reading center setting with sophisticated training of readers.

The wide applicability of our deep learning framework is another strength. While other works for iRORA/cRORA detection in OCT tested their model performance on data from one specific device [282, 283], we employed a DA approach to allow the generalization to data from other devices, without the need for manually annotated training data from these devices. This generalization was tested on data from two main OCT devices, making our algorithm relevant for future multicenter studies where data from different imaging devices are available. The availability of reliable AI models transferable to various OCT devices can further help to identify patients at high risk for disease progression in need of more frequent follow-up visits in our daily routine as well as potential candidates for upcoming interventional study trials in iAMD. Moreover, the biomarker segmentation model [68], was trained on more AMD disease-related features than the ones currently used. This indicates the potential of our algorithm to be directly applied to other retinal OCT applications, such as fluid segmentation for exudative AMD patients, ellipsoid zone loss detection for retinitis pigmentosa, or the segmentation of intraretinal cysts for X-linked retinoschisis. Without any further model development, our existing algorithm may already generalize well to OCT data with these pathologies from the various devices used in this work, due to the DA approach that we employed. Future work could focus on further validating this broader applicability across various retinal conditions.

This study has various limitations. (1) While the MACUSTAR study consists of both a cross-sectional and longitudinal part, we only used the data from the cross-sectional part, as we did not have access to the longitudinal part yet [266]. Future work should aim to incorporate this longitudinal data to evaluate our approach further. Furthermore, this work could be extended to predict disease progression with the use of AI. (2) The number of cases with iRORA/cRORA was relatively small. Specifically, based on manual annotations, there were 22 OCTs with iRORA/cRORA in the Spectralis subset and 11 in the ZEISS subset. In future analyses, more data should be included, including both the longitudinal and additional cross-sectional data beyond the iAMD subset from the MACUSTAR study, or even datasets from other studies. This may, for example, lead to better estimates of the performance estimates in general and statistically significant results when measuring the effect of including DA in the deep learning framework. (3) We employed only a single modality

and did not incorporate *en face* imaging from other modalities, which could have allowed for a more accurate detection of iRORA/cRORA. (4) RPE tear is an exclusion criterion for iRORA/cRORA [62], which was not taken into account by our deep learning model. (5) The DA approach we used in the current work only took differences related to the appearance of individual B-scans, such as the noise pattern, into account. However, there is also a lot of variability in the spacing between B-scans across OCT devices, and vertical registration strategies in the software of these devices. These differences may also influence biomarker quantification, especially given that we used a 3D segmentation model, but they were not addressed by the DA approach in this work. (6) In some cases, the optic disc fell in the 6 × 6 mm *en face* square we took around the fovea during pre-processing, which sometimes led to our model wrongly indicating iRORA/cRORA at the location of the optic disc. Future iterations of our model should incorporate a mechanism against this, *e.g.,* by introducing such samples in the training set.

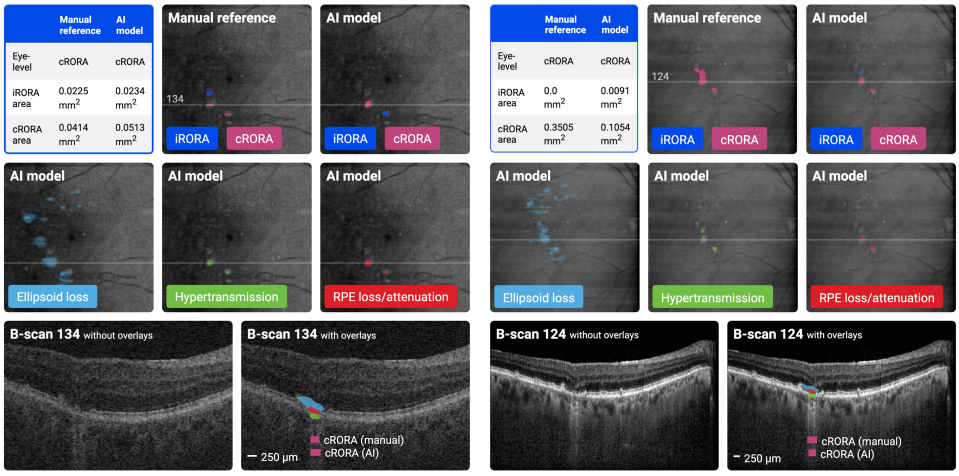
In conclusion, we present a robust deep learning algorithm to automatically detect iRORA/cRORA from SD-OCT. This algorithm is generalizable across various OCT devices, while manually annotated development data was only required from one of these devices. Applications of this model may enable precise automated tracking of AMD-related lesion changes in numbers, size, area, and retinal position. This is of major importance for monitoring iAMD disease progression, particularly given the increasing amount of multimodal retinal imaging data becoming available in studies.

Acknowledgements

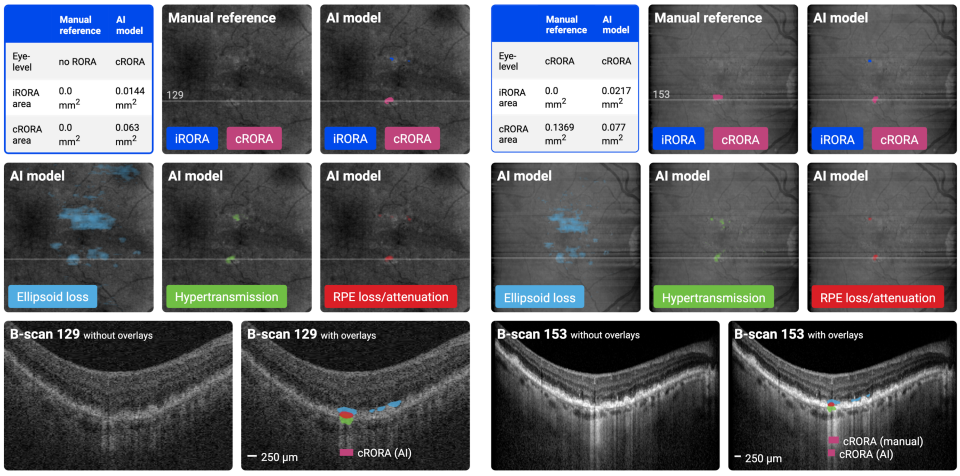
This work was supported by the FEMHabil Program, Faculty of Medicine, University of Bonn to MS. This project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement No. 116076. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations (EFPIA). Supported in part by an Unrestricted Grant from Research to Prevent Blindness, New York, NY, to the Department of Ophthalmology & Visual Sciences, University of Utah.

MACUSTAR consortium members: H. Agostini, I. Aires, L. Altay, R. Atia, F. Bandello, P. G. Basile, J. Batuca, C. Behning, M. Belmouhand, M. Berger, A. Binns, C. J. F. Boon, M. Böttger, A. R. Branco, J. E. Brazier, C. Carapezzi, J. Carlton, A. Carneiro, A. Charil, R. Coimbra, D. Cosette, M. Cozzi, D. P. Crabb, J. Cunha-Vaz, C. Dahlke, H. Dunbar, R. P. Finger, E. Fletcher, M. Gutfleisch, F. Hartgers, B. Higgins, J. Hildebrandt, E. Höck, R. Hogg, F. G. Holz, C. B. Hoyng, A. Kilani, J. Krätzschmar, L. Kühlewein, M. Larsen, S. Leal, Y. T. E. Lechanteur, D. Lu, U. F. O. Luh-

mann, A. Lüning, N. Manivannan, I. Marques, C. Martinho, K. P. Moll, G. Montesano, Z. Mulyukov, M. Paques, B. Parodi, M. Parravano, S. Penas, T. Peters, T. Peto, S. Poor, S. Priglinger, R. Ramamirtham, D. Rowen, G. S. Rubin, J. Sahel, C. Sánchez, O. Sander, M. Saßmannshausen, M. Schmid, S. Schmitz-Valckenberg, J. Siedlecki, R. Silva, E. Souied, G. Staurenghi, J. Tavares, D. J. Taylor, J. H. Terheyden, A. Tufail, P. Valmaggia, M. Varano, A. Wolf, and N. Zakaria.



(a)



(b)

Figure 6.6: Qualitative comparison of manual reference and AI model outputs between OCTs from the two devices. Each subfigure shows the ZEISS OCT first, followed by the Spectralis OCT, both taken from the same eye during the same visit. The B-scans presented below in each subfigure approximately correspond to the same location in the retina for both OCT devices. Each grid in the subfigures shows a table in the top left, containing the eye-level decisions and the *en face* iRORA/cRORA areas, calculated from the segmentations, for both the manual reference and the AI model. The middle and right cells of the first rows show the iRORA/cRORA segmentations from the manual reference and the AI model, respectively, overlaid on the *en face* projection of the OCT volume. The middle rows show the segmentations from the AI model for the three underlying granular biomarkers of iRORA/cRORA, overlaid on the *en face* projection of the OCT volume. Each *en face* image shows a white horizontal line, which corresponds to the B-scan shown in the last row. The last rows show the AI model segmentation of the granular biomarkers. If iRORA/cRORA is detected, the last rows also show horizontal lines labeled as either "iRORA/cRORA (AI)" for the AI model or "iRORA/cRORA (manual)" for the manual reference. (Caption continues on next page.)

Figure 6.6: (Caption continuation from previous page.) **(a)** An example where the outcomes are consistent between devices. Both the AI model and the manual reference agree on the presence of a cRORA lesion in the ZEISS and Spectralis OCT. **(b)** An example with a discrepancy between the manual reference for the two devices. The output of the AI model is more consistent regarding the detected iRORA and cRORA lesions in the OCTs from both devices. The manual reader likely missed the lesion in the ZEISS OCT due to human error, given the evident presence of all criteria for cRORA. A sub-analysis of all ZEISS OCTs with a discrepancy between the manual and AI eye-level decisions, similar to what was done for the Spectralis OCTs in this study, would have likely led to the manual delineation of this lesion.

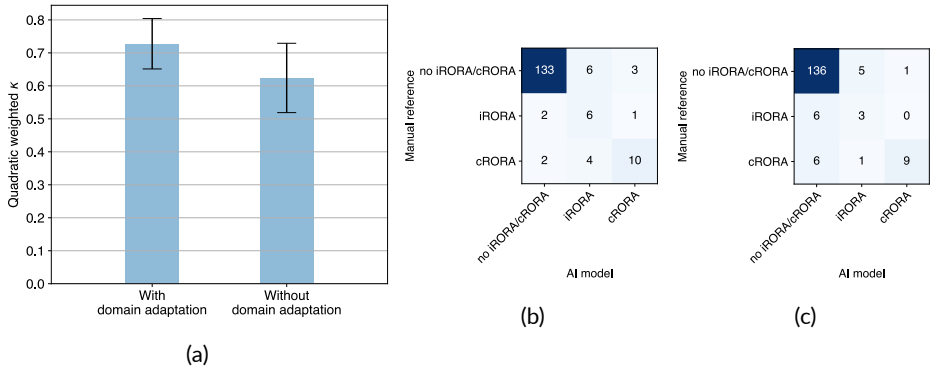


Figure 6.7: The effect of domain adaptation (DA) on eye level None/iRORA/cRORA classification in the Spectralis dataset. **(a)** The effect on the quadratic weighted kappa score (between the manual annotations and the AI model). The error bars are standard deviations from non-parametric bootstrapping with 1,000 iterations. **(b)** Confusion matrix with DA. (c) Confusion matrix without DA.

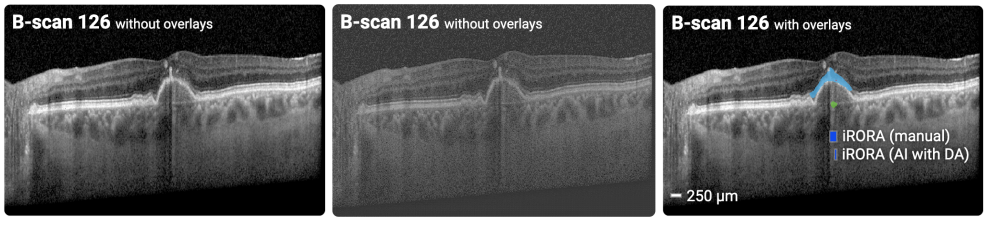


Figure 6.8: Example where the AI model with domain adaptation (DA) detected an iRORA lesion that was not picked up without DA.

Chapter appendix

6.A Additional results

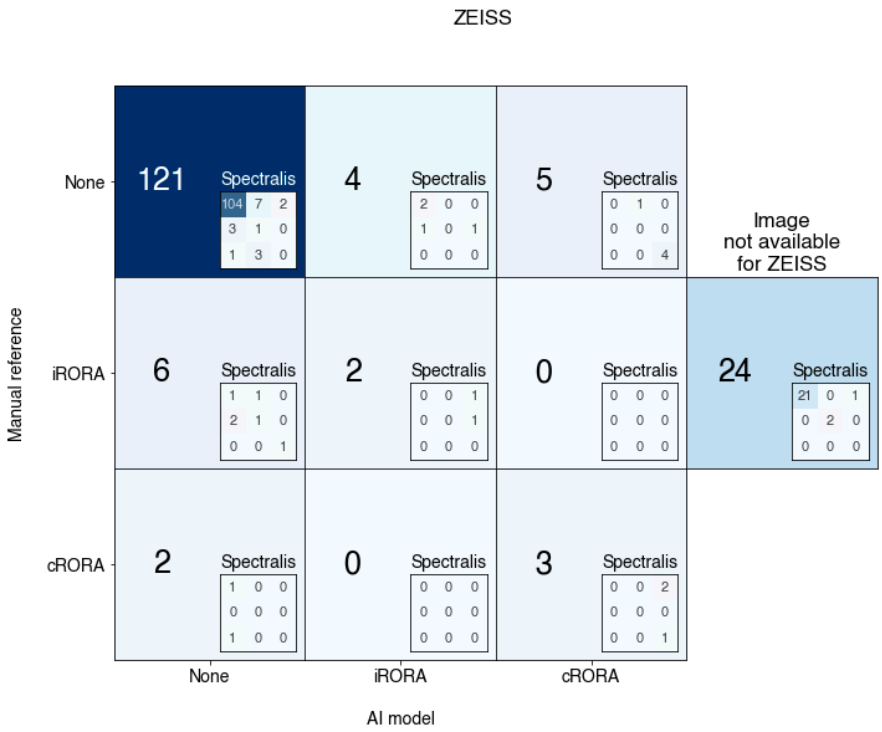


Figure 6.A.1: Confusion matrix visualizing the eye level agreement between the AI model and manual reference for both ZEISS and Spectralis OCT data. For Spectralis, the updated labels were used. The numbers printed in larger fonts represent the count of ZEISS OCTs, with a smaller confusion matrix for corresponding Spectralis OCTs provided in smaller print. For enhanced readability, axis labels on these smaller confusion matrices were omitted but mirror those of the main matrix. For example, the cell in the second row and column signifies the AI model's agreement with the manual reference on the classification of iRORA for ZEISS OCTs. However, when evaluating their corresponding Spectralis OCTs, both the AI model and manual reference classified them as cRORA.

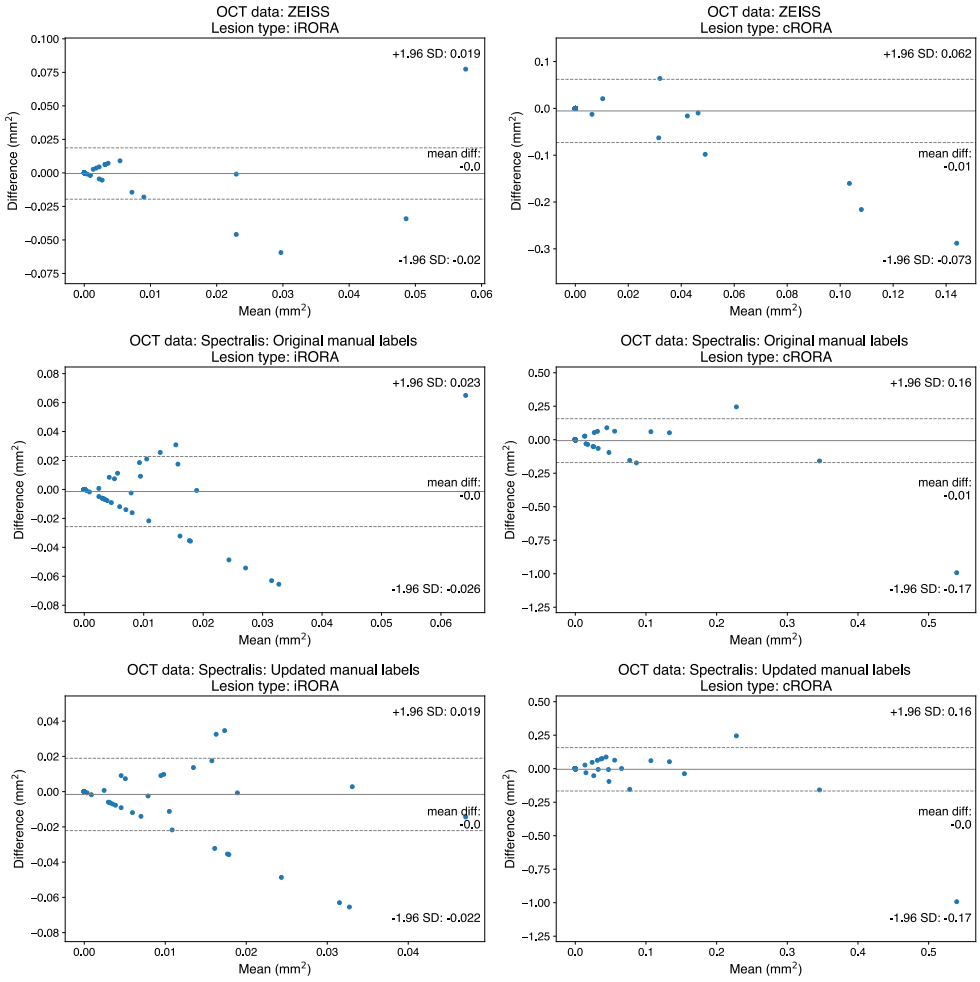


Figure 6.A.2: Bland-Altman plots comparing the AI model and manual reference for iRORA and cRORA en face areas. The results are shown separately for the ZEISS and Spectralis OCTs. The Spectralis results are calculated using the updated manual labels. The differences were calculated as: "area of the manual annotations - area of the AI segmentation".

Table 6.A.1: Validation results of the domain adaptation (DA) model on the RETOUCH test set. The results are reported as the Dice scores, denoted as mean (standard deviation).

Trained on	Evaluated on	Intraretinal fluid	Subretinal fluid	Pigment epithelial detachment
Topcon	Spectralis	0.71 (0.114)	0.37 (0.31)	0.61 (0.30)
Topcon	Spectralis (Topcon appearance generated using DA)	0.74 (0.10)	0.42 (0.26)	0.64 (0.32)

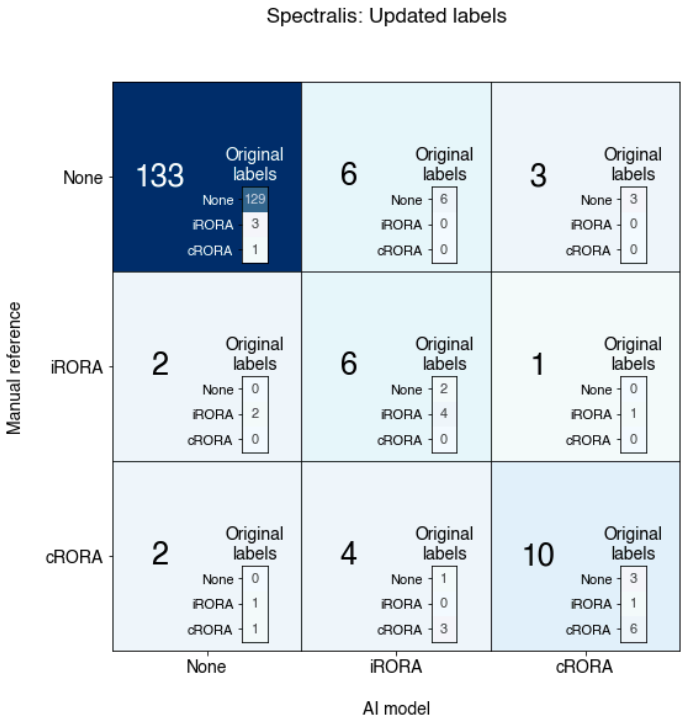


Figure 6.A.3: Confusion matrix of the updated manual labels in the Spectralis OCTs. In each cell, it is indicated in a smaller font what the original labels were for the OCTs within that cell.

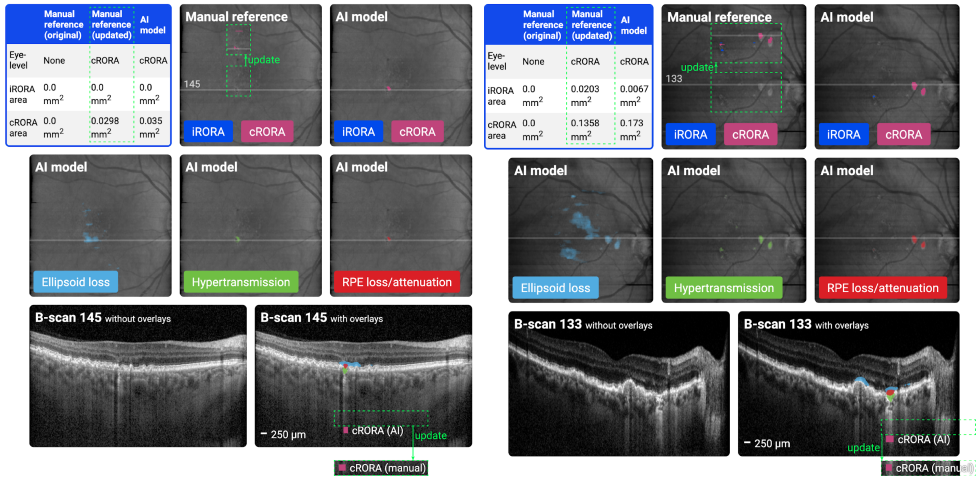


Figure 6.A.4: Instances where manual labels in Spectralis OCTs were revised based on discrepancies between the AI model outputs and the original manual reference at the eye level. Each subfigure shows a table in the top left, containing the eye-level decisions and the en face iRORA/cRORA areas, calculated from the segmentations, for both the manual reference and the AI model. The middle and right cells of the first rows show the iRORA/cRORA segmentations from the manual reference and the AI model, respectively, overlaid on the en face projection of the OCT volume. The middle rows show the segmentations from the AI model for the three underlying granular biomarkers of iRORA/cRORA, overlaid on the en face projection of the OCT volume. Each en face image shows a white horizontal line, which corresponds to the B-scan shown in the last row. The last rows show the AI model segmentation of the granular biomarkers. If iRORA/cRORA is detected, the last rows also show horizontal lines labeled as either “iRORA/cRORA (AI)” for the AI model or “iRORA/cRORA (manual)” for the manual reference. The green dotted boxes indicate changes resulting from these revisions.

"What I cannot create, I do not understand."

Richard Feynman

7

Standardizing OCT through *en face* conditioned 3D diffusion models

Authors: Coen de Vente, Mohammad M. Islam, Philippe Valmaggia, Carel B. Hoyng, Adnan Tufail, Clara I. Sánchez on behalf of the MACUSTAR Consortium[†] ([†]see chapter acknowledgements)

Original title: Conditioning 3D Diffusion Models with 2D Images: Towards Standardized OCT Volumes through En Face-Informed Super-Resolution

Published in: NeurIPS Workshop on GenAI for Health: Potential, Trust and Policy Compliance (2024)

Abstract

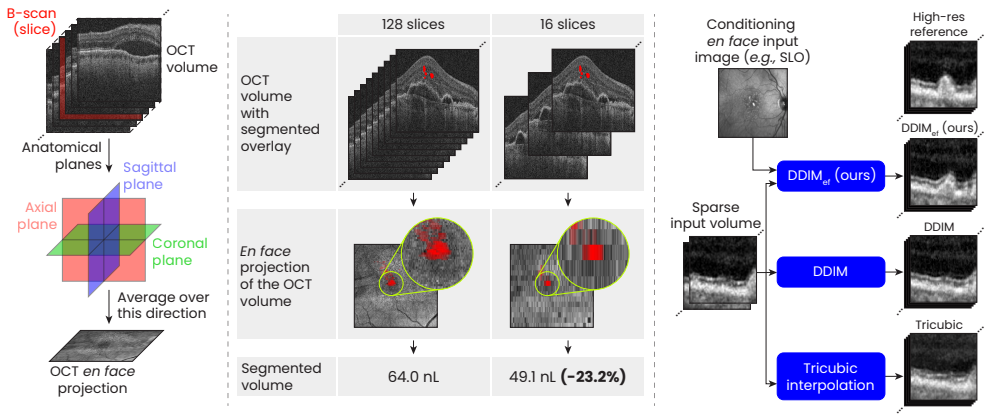
High anisotropy in volumetric medical images can lead to the inconsistent quantification of anatomical and pathological structures. Particularly in optical coherence tomography (OCT), slice spacing can substantially vary across and within datasets, studies, and clinical practices. We propose to standardize OCT volumes to less anisotropic volumes by conditioning 3D diffusion models with *en face* scanning laser ophthalmoscopy (SLO) imaging data, a 2D modality already commonly available in clinical practice. We trained and evaluated on data from the multicenter and multimodal MACUSTAR study. While upsampling the number of slices by a factor of 8, our method outperforms tricubic interpolation and diffusion models without *en face* conditioning in terms of perceptual similarity metrics. Qualitative results demonstrate improved coherence and structural similarity. Our approach allows for better informed generative decisions, potentially reducing hallucinations. We hope this work will provide the next step towards standardized high-quality volumetric imaging, enabling more consistent quantifications.

7.1 Introduction

Volumetric medical images can be highly anisotropic, *i.e.,* having high-resolution slices in one anatomical plane but poor through-plane resolution. This has been shown to lead to imprecise volume and shape measurements of structures of interest [294], potentially resulting in wrong diagnoses and severe negative clinical implications.

A prominent example of a modality often affected by this is optical coherence tomography (OCT). OCT is commonly acquired as a raster, where multiple line scans generate B-scans (slices), and multiple slices generate a volume (see Fig. 7.1a). The spacing between slices can vary substantially between OCT devices and imaging protocols [66,253,295,296]. This can hamper consistent biomarker quantification [69, 297–300]. Fig. 7.1b illustrates this issue by demonstrating fluid volume estimations in a single retinal OCT volume for various slice spacings. We can observe a 23.2% drop in estimated fluid volume when the slice spacing increases by a factor of 8.

A possible solution to these imprecise measurements is to standardize volumes with low slice density to high-resolution data through reliable super-resolution methods. Several super-resolution approaches have been proposed for OCT [301–304] but they all aim to improve the resolution within individual B-scans. Approaches to reduce anisotropy have been proposed for other volumetric medical images, such as computed tomography (CT) [305] and magnetic resonance imaging (MRI) [306]. These methods use deep learning



(a) Medical terminology. (b) Problem statement. (c) Our solution versus baselines.

Figure 7.1: Overview of our proposed approach and background. **(a)** The medical terminology related to OCT volumes. **(b)** Low slice densities in volumetric images can lead to inaccurate quantifications. **(c)** We propose to upsample the number of slices by conditioning a diffusion model with *en face* images. In our experiments, we use SLO as conditioning *en face* data but our method could be extended to other types. In the shown example, our model is the only approach that correctly generates the druse (the bright bump in the retina).

models to upsample the number of slices based solely on low-resolution input data during inference.

A major drawback of these works is their lack of knowledge about anatomical and pathological structures that fall between two adjacent slices. This can lead to hallucinating models that generate incorrect biological features, potentially resulting in misdiagnoses or otherwise harmful clinical outcomes. We hypothesize that including information about regions between slices as input to a super-resolution model helps make better-informed generative decisions that correctly reflect the biological truth.

Therefore, we propose a method based on 3D diffusion models to increase slice density by utilizing additional imaging modalities as conditioning information (see Fig. 7.1c). We use diffusion models, as they have been shown to outperform other popular generative models such as generative adversarial networks at generating high-quality images [307], super-resolution [308], and leveraging multimodal data as conditioning information [12]. These capabilities align well with the objectives of our study.

We evaluate our developed method on OCT data while conditioning on scanning laser ophthalmoscopy (SLO) fundus images. SLO is a 2D *en face* (*i.e.,* parallel to the coronal plane) imaging modality that is commonly acquired alongside OCT scans. OCT devices internally use SLO images as a reference to position the OCT acquisition at the desired anatomical location [55]. Our method can be extended to include other modalities, such as color fundus photography (CFP) and fundus autofluorescence (FAF), potentially resulting in even better-informed models.

We hope this approach is a valuable step towards more isotropic, high-quality, and standardized volumetric imaging, allowing for more consistent biological measurements and diagnoses in the future.

7.2 Methods

Our model is a 3D diffusion model trained to generate high-resolution volumes by upsampling the number of slices (*i.e.,* increasing the B-scan density in OCT). It is conditioned on both *en face* imaging data and the low-resolution counterpart of the high-resolution target (see Fig. 7.2). We provide a brief introduction to diffusion models and their associated symbols in Section 7.2.1. In Section 7.2.2, we describe how we adapt diffusion models for *en face* conditioned super-resolution. In short, the low-resolution image is concatenated with a reshaped *en face* image along the channel dimension, which we subsequently input as conditional information to the denoising model. The sampling process, including the use of Denoising Diffusion Implicit Model (DDIM) [309] sampling, overlapping patches en-

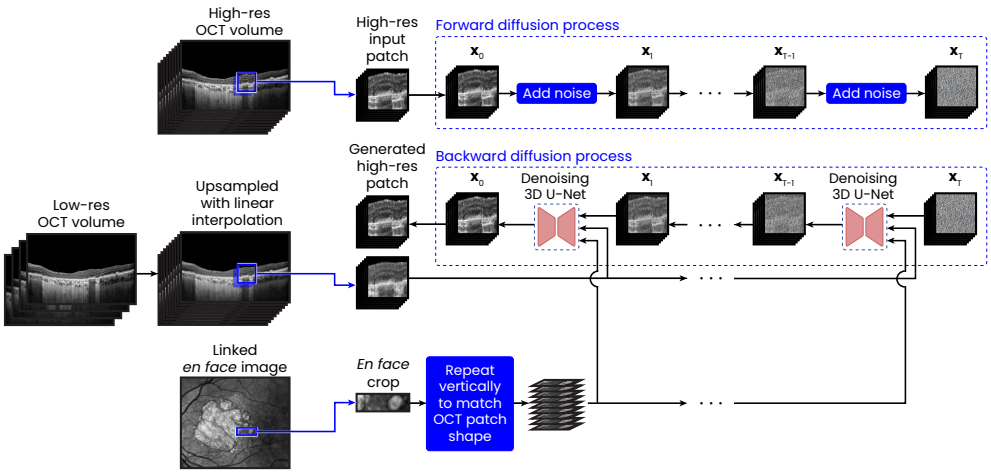


Figure 7.2: *En face*-conditioned diffusion model overview.

abled by RePaint [310], and classifier-free guidance (CFG) [311], is detailed in Section 7.2.3. Finally, we present the network architecture and implementation details in Section 7.2.4.

7.2.1 Diffusion models

Diffusion models are generative models consisting of a forward diffusion process and a backward diffusion process [312]. In the forward diffusion process, over many timesteps T, more and more noise is gradually added to an input image $\mathbf{x}_0$, resulting in noisy images $\mathbf{x}_1, \ldots, \mathbf{x}_T$. This process q can be formulated with a variance schedule $\beta_1, \ldots, \beta_T$ as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{7.1}$$

As shown by Ho *et al.* [312], we can directly obtain $\mathbf{x}_t$ given $\mathbf{x}_0$ using the following equation:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{7.2}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

In the backward diffusion process p_θ, $\mathbf{x}_{t-1}$ is predicted for any $t \in \{1, \ldots, T\}$ by denoising $\mathbf{x}_t$ using a trained denoising model, optimized by model parameters θ. This model generally uses some variant of the U-Net [9] architecture (see Section 7.2.4 for the implementation we use). Following Salimans *et al.* [313], our trained denoising model does not predict the noise $\boldsymbol{\epsilon}$ or image $\mathbf{x}_0$ directly, but uses $\mathbf{v}$-prediction parameterization as this prevents in-

tensity shifting artifacts in super-resolution models [314], where $\mathbf{v}_t := \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\mathbf{x}_0$. We use the mean squared error (MSE) loss to train our denoising model v_θ:

$$\mathcal{L} := ||v_\theta(\mathbf{x}_t, t) - \mathbf{v}_t||_2^2. \tag{7.3}$$

7.2.2 *En face* conditioned super-resolution

To enable the use of diffusion models for *en face* conditioned super-resolution, we combine the approach for generating high-resolution from low-resolution images from SR3 [308] with our proposed method of conditioning the denoising model with *en face* information. In this paper, we use SLO images but our approach could be extended to other *en face* data.

Specifically, we adapt the input to the conditional denoising model as follows:

$$\hat{\mathbf{v}}_t = v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}), \tag{7.4}$$

where $\hat{\mathbf{v}}_t$ is the output of the denoising model, $\mathbf{x}_t$ is the noisy image at timestep t, $\mathbf{x}_{LR}$ is the corresponding low-resolution image upsampled to match the dimensions of $\mathbf{x}_t$ using linear interpolation, and $\mathbf{x}_{enface}$ is the *en face* image.

$\mathbf{x}_{LR}$ and $\mathbf{x}_{enface}$ are fed similarly into the denoising model. Following SR3 [308], we concatenate the noisy input $\mathbf{x}_t$ with the low-resolution image $\mathbf{x}_{LR}$ along the channel dimension. Since *en face* images generally do not align with their corresponding OCT scans in the *en face* plane, registration of these two images is required (see Section 7.A). This registration step results in spatial correspondence and ensures the *en face* image has the same shape in the *en face* plane as the OCT volume. To reach an image with the same 3D shape as $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, we repeat the registered 2D $\mathbf{x}_{enface}$ image in the y-direction H times, where H is the height of $\mathbf{x}_t$. This 3D tensor is concatenated together with $\mathbf{x}_t$ and $\mathbf{x}_{LR}$ along the channel dimension. We subsequently input the resulting tensor into the model.

Due to computational limitations, we work with image crops for $\mathbf{x}_t$, $\mathbf{x}_{LR}$, and $\mathbf{x}_{enface}$. These crops all correspond in terms of their location and size. No noise is added to $\mathbf{x}_{LR}$ and $\mathbf{x}_{enface}$ for any timestep t.

7.2.3 Sampling process

During sampling, there are a few key distinctions in the processing pipeline compared to the backward diffusion process during training. Firstly, we use DDIM [309] sampling, which allows for accelerated sampling by reducing the number of timesteps while sampling.

Secondly, we train our denoising model with patches, but we are interested in generating full high-resolution volumes during sampling. To prevent artifacts near the borders of patches, we use overlapping patches. We use RePaint [310] to facilitate this overlapping strategy, an image inpainting approach for diffusion models (see Section 7.B for more details).

Thirdly, to minimize image artifacts showing an overall intensity in the generated slices that is different from the overall intensity in the slices already existing in the low-resolution volume, we implemented a post-processing normalization step. In this step, we normalized the mean and standard deviation of the intensities in the generated slices to match those in the slices that already existed in the low-resolution volume. This normalization step was performed separately for each OCT volume.

Lastly, to influence how much the denoising model uses the *en face* information for its generative decisions, we employ CFG [311]. In CFG, during training, the conditional information is dropped for a random number of samples in each batch with some probability p_{uncond}. In practice, when the conditional information is dropped, we feed an image with all pixels set to zero. This results in a jointly trained conditional denoising model $v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface})$ and unconditional denoising model $v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR})$. During sampling, we can then linearly combine the conditional and unconditional model predictions using a guidance scale hyperparameter w:

$$\tilde{v}_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}) = (1-w)v_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}) + wv_\theta(\mathbf{x}_t, t, \mathbf{x}_{LR}, \mathbf{x}_{enface}). \tag{7.5}$$

7.2.4 Network architecture and implementation details

For our denoising network, we use the U-Net architecture described by Pinaya *et al.* [315], in which the timestep embedding is integrated into the model via residual connections. The U-Net uses 3D convolutions and has a depth of four U-Net levels with 32, 64, 128, and 256 channels, respectively, in each level with two residual blocks per level. We use self-attention at the deepest U-Net level with a single attention head. We train using the Adam optimizer with a learning rate of 5×10^{-5}, a batch size of 16, for 20 000 epochs. All images were normalized between -1 and 1 before cropping.

In our experiments, the number of timesteps for the diffusion process was $T = 1000$ during training with a SCALED LINEAR [315] schedule and a β_t range of 0.0005 to 0.0195. We employed DDIM sampling with 100 timesteps, resulting in a 10$\times$ time efficiency improvement. We do not perform any resampling steps when inpainting using RePaint [310] to allow for faster sampling. We use the MONAI generative AI framework [315] for imple-

menting our diffusion model.

During training, we use 3D crops of size $128 \times 128 \times 16$ for $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, and a 2D crop of size $128 \times 1 \times 16$ for $\mathbf{x}_{enface}$. During sampling, we use a crop size of $496 \times 496 \times 16$ for $\mathbf{x}_t$ and $\mathbf{x}_{LR}$, and a 2D crop of size $496 \times 1 \times 16$ for $\mathbf{x}_{enface}$. When sampling full high-resolution volumes, we used an overlap of 25%, 25%, and 50% for the x-, y-, and z-direction, respectively. During training, we prepared our crops such that each 5$^{\text{th}}$ and each 13$^{\text{th}}$ slice in $\mathbf{x}_{LR}$ were identical to the 5$^{\text{th}}$ and 13$^{\text{th}}$ slice in $\mathbf{x}_0$, respectively. The other slices were interpreted using linear interpolation.

7.2.5 Data

For training and evaluating our diffusion models, we used OCT volumes and corresponding SLO images from the MACUSTAR study [266]. MACUSTAR is a clinical study on age-related macular degeneration (AMD) that is carried out across 20 sites in 7 European countries. The dataset from this study contains data from patients with varying disease severities (no, early, intermediate, and advanced AMD).

We used the patient data from the cross-sectional part of the MACUSTAR study and only included Heidelberg Spectralis OCTs with at least 237 B-scans. This resulted in a total set of 302 patients. We randomly split this set of patients in 181 (60%), 60 (20%), and 61 (20%) patients for training, validation, and testing, respectively. As multiple OCT volumes were available for each patient (from both eyes and either one, two, or three visits in the cross-sectional study) this resulted in 721 and 236 OCT volumes for the training and validation set, respectively. For the test set, we only used the OCT volume from the study eye, defined for the MACUSTAR study, from the first visit, resulting in 61 OCT volumes. More details regarding the dataset and pre-processing can be found in Appendix 7.A.

7.3 Experiments

We evaluate our approach for the task of upsampling the number of slices in the image volume with a factor of 8. For the sake of simplicity, when generating full volumes as described in Section 7.2.3, we drop the last slice in the OCT volume in the test set, resulting in OCT volumes with 240 instead of 241 slices. Hence, we upsample low-resolution volumes with 30 slices to high-resolution volumes with 240 slices. The resolution of individual slices was not changed.

We refer to our proposed diffusion model with *en face* conditioning and CFG with guidance scale $w = 2$ as DDIM$_{\text{ef}}$. To measure the effect of *en face* conditioning and CFG, we compare DDIM$_{\text{ef}}$ with the two models: DDIM$_{\text{ef}}$ (no CFG), and DDIM. DDIM is the proposed

approach with *en face* conditioning turned off during sampling. *En face* conditioning is turned off by feeding an image with all pixels set to zero as the conditional image, which is also done during training with a probability of p_{uncond} to enable CFG (see Section 7.2.3). The proposed model with *en face* conditioning, but without CFG, will be referred to as DDIM_{ef} (no CFG). Furthermore, we compare these methods with the more traditional upsampling method of tricubic interpolation.

7.3.1 Evaluation

We report the classical image similarity metrics MSE, SSIM [316], and PSNR, computed between the $8\times$ upsampled low-resolution images using DDIMs and tricubic interpolation, and the high-resolution reference images. As noted by Saharia *et al.* [308], these classical metrics may not be optimal for evaluating super-resolution methods, as they were shown to correlate poorly with human perception and heavily penalize synthetic high-frequency details that deviate from the reference, favoring than blurry images instead.

Therefore, we also evaluate with Learned Perceptual Image Patch Similarity (LPIPS) [317]. We used the LPIPS metric implementation provided by the authors of the original LPIPS paper [317] for their metric based on an ImageNet pre-trained AlexNet [4]. Since this evaluation method was developed for 2D images, we modified it for 3D data by calculating the LPIPS metric on all 2D slices in the axial, coronal and sagittal planes in the volume, resulting in the metrics $\text{LPIPS}_{\text{axi}}$, $\text{LPIPS}_{\text{cor}}$, and $\text{LPIPS}_{\text{sag}}$, respectively. We also report $\text{LPIPS}_{\text{2.5D}}$, which is the average of these three metrics. This approach for 3D data is available in the publicly available implementation from the MONAI generative AI framework [315]. Additionally, we calculated $\text{LPIPS}_{\text{efproj}}$, which compares two OCT *en face* projections generated by averaging all columns in each slice of the volume, resulting in a 2D *en face* image.

Even though LPIPS may be considered a more suitable evaluation approach than the classical evaluation metrics, it is not guaranteed to consider the structures of interest enough, since the underlying model was only trained on natural images. Therefore, we also present qualitative examples.

7.4 Results

The first three rows of Table 7.1 shows the classical image similarity metrics MSE, SSIM [316], and PSNR, computed between the $8\times$ upsampled low-resolution images using DDIMs and tricubic interpolation, and the high-resolution reference images. The perceptual metrics based on LPIPS are shown in the last five rows of Table 7.1.

Table 7.1: Classical image similarity metrics (MSE, SSIM, and PSNR) and perceptual metrics (all LPIPS variants) calculated on the test set between the original high-resolution OCT volumes and low-resolution images that were 8× upsampled in the slice-direction using tricubic interpolation and our proposed DDIM methods. Results are presented as the mean ± standard deviation over all OCT volumes. Bolded values indicate the best values in terms of the mean performance.

	Tricubic	DDIM	DDIM_{ef} (no CFG)	DDIM_{ef}
MSE ↓	**0.006** ± **0.002**	**0.006** ± **0.003**	**0.006** ± **0.003**	**0.006** ± **0.003**
SSIM ↑	**0.451** ± **0.116**	0.444 ± 0.107	0.447 ± 0.107	0.447 ± 0.107
PSNR (dB) ↑	22.472 ± 1.418	22.401 ± 1.644	**22.495** ± **1.673**	22.450 ± 1.683
$\text{LPIPS}_{\text{axi}}$ ↓	**0.120** ± **0.027**	0.138 ± 0.030	0.138 ± 0.030	0.141 ± 0.031
$\text{LPIPS}_{\text{cor}}$ ↓	0.548 ± 0.103	**0.158** ± **0.047**	**0.158** ± **0.048**	0.162 ± 0.050
$\text{LPIPS}_{\text{sag}}$ ↓	0.540 ± 0.088	**0.144** ± **0.049**	**0.144** ± **0.049**	0.147 ± 0.050
$\text{LPIPS}_{\text{2.5D}}$ ↓	0.403 ± 0.072	**0.147** ± **0.041**	**0.147** ± **0.042**	0.150 ± 0.043
$\text{LPIPS}_{\text{efproj}}$ ↓	0.231 ± 0.055	0.063 ± 0.039	**0.060** ± **0.039**	0.064 ± 0.039

For the proposed method, ablated methods and the tricubic interpolation method, we present several figures to illustrate the difference in structural similarity to the high-resolution reference images, sharpness, and coherence within the generated volumes. In Fig. 7.3 and 7.C.1, we aim to point out these aspects using the *en face* projections, allowing one to observe the overall structure of the full generated volumes. Fig. 7.4 and Fig. 7.C.2 show additional examples of these *en face* projections while highlighting relevant regions, alongside the corresponding image crops from the underlying OCT volume. Further examples are presented in Fig. 7.C.3, which shows 3D renders of generated and reference volumes, and in Fig. 7.C.4, which shows crops of consecutive B-scans with difference maps between generated and reference images. Fig. 7.C.5 shows generation examples and difference maps for several randomly picked image crops. The effect of increasing the guidance scale w from CFG is illustrated in Fig. 7.5. Specifically, a case of an example OCT crop is shown with several drusen that are also visible in the SLO image.

Model training took approximately 9 days on an NVIDIA A100 GPU. Sampling a full volume with 240 B-scans of 768 × 496 pixels, the most common size in the test set, resulted in 58 patches. On the aforementioned GPU type, for DDIM_{ef}, sampling a whole volume took approximately 46 minutes. For DDIM and DDIM_{ef} (no CFG), this sampling time was about half (approximately 23 minutes), as CFG doubles the number of required forward passes of the denoising model.

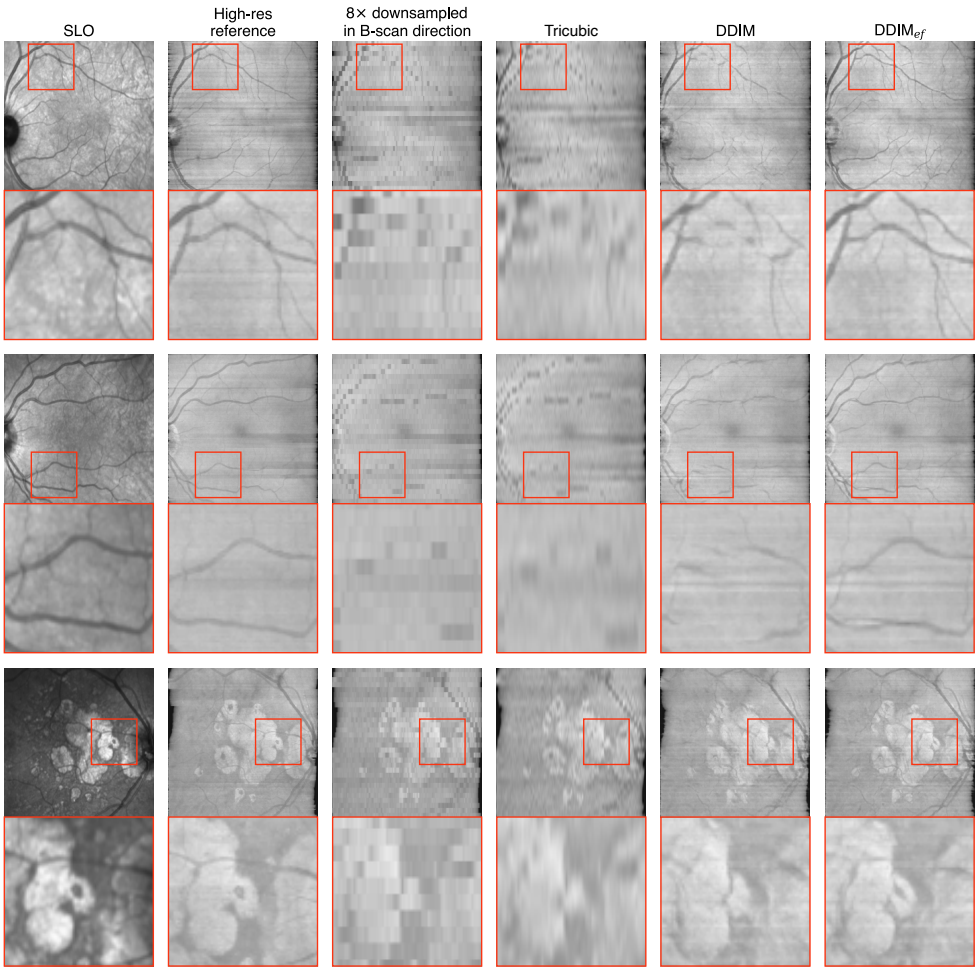


Figure 7.3: Examples of *en face* projections of the OCT volumes generated using tricubic interpolation, the unconditional diffusion model DDIM, and our proposed *en face* conditioned diffusion model DDIM$_{ef}$. These three projections are shown in the last three columns and were all generated from the 8 $\times$ downsampled (in the B-scan direction) volume as input, which is shown in the third column. The first and second rows show the corresponding high-resolution (high-res) reference and scanning laser ophthalmoscopy (SLO) image, respectively. A separate example from a different test set patient is shown in each row. The top images in each row show the full image. Zoomed-in versions of the image crops (red boxes) are shown at the bottom of each row.

7.5 Discussion

We addressed the large variability in anisotropy across OCT scans, which can lead to inconsistent quantifications. We propose a super-resolution approach that uses SLO images to condition 3D super-resolution diffusion models, aiming for better informed image generations that are closer the biological truth. SLO is commonly already acquired alongside

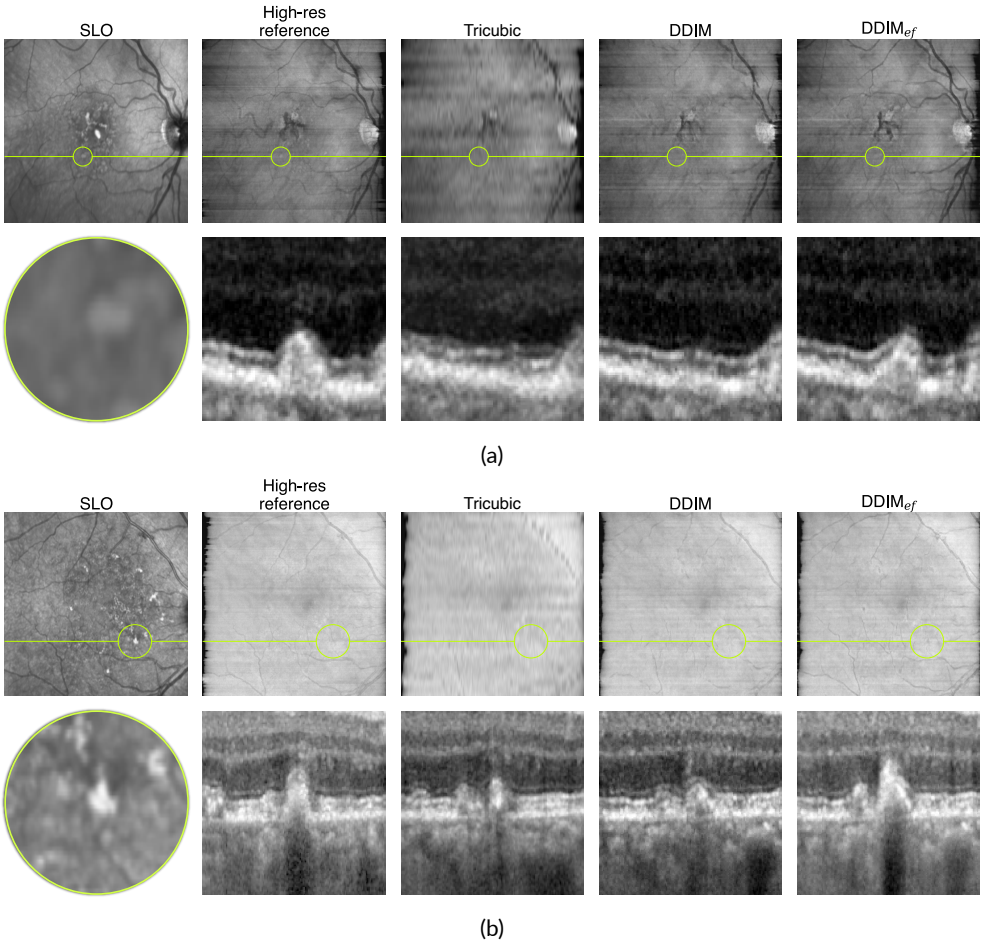


Figure 7.4: Examples showing the effect of using DDIM$_{ef}$, compared to leaving out the *en face* information (DDIM), and tricubic interpolation. In the top rows, the SLO image is shown on the left, followed by the *en face* projections of each OCT volume. The bottom rows show zoomed-in crops. The crop locations are indicated in the top row with a lime circle. The lime horizontal line corresponds to the B-scan location from the shown crops. **(a)** DDIM$_{ef}$ reconstructs a druse that is present in the high-resolution reference, while it is missing in the OCT crops from tricubic interpolation and DDIM. The druse also seems to be subtly visible in the SLO image. **(b)** The large lesion in the center of the OCT crops are more similar to the reference for DDIM$_{ef}$ than for tricubic interpolation and DDIM. It is also visible in the SLO images as a hyper-intense lesion.

OCT scans, ensuring our method often will not require any additional data beyond what is already available in clinical practice. Furthermore, SLO acquisition is relatively fast, while OCT acquisition time increases with every additional line to be acquired, potentially speeding up the overall acquisition process.

Our qualitative results indicate that our approach can upsample the number of B-scans in OCT volumes by a factor of 8 while improving similarity to high-resolution reference

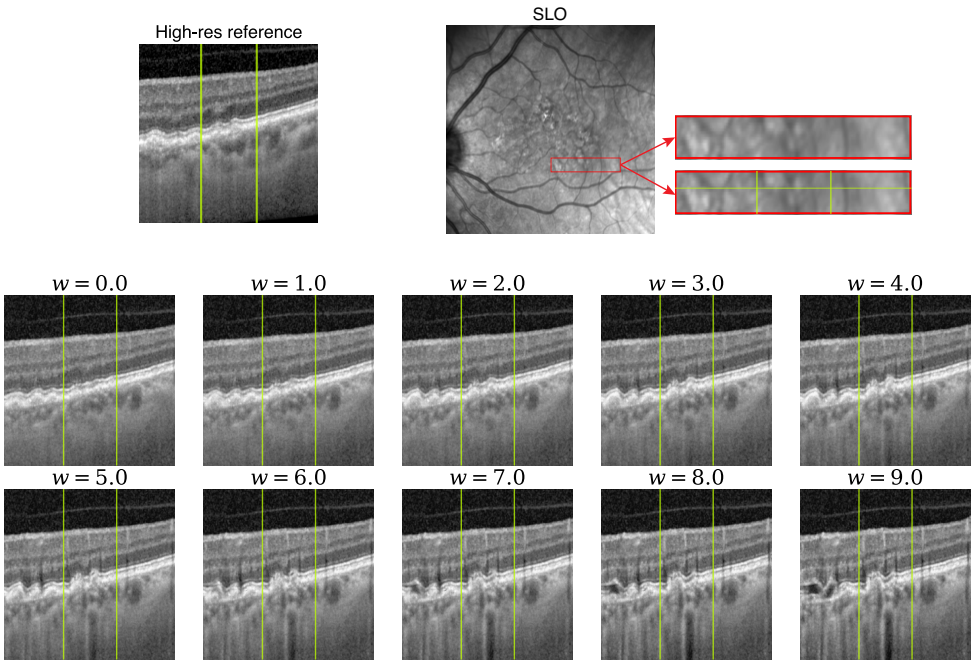


Figure 7.5: The effect of CFG and increasing the guidance scale w. In the top left, the high-resolution reference is shown. In the top right, the *en face* SLO image is shown, in which the crop location is indicated with a red box. The vertical lime lines correspond to the same physical locations throughout the figure. The horizontal lime line in the top right indicates the B-scan location of the shown OCT crops. The bottom two rows show the results from our *en face* conditioned diffusion model with CFG for various guidance scales. $w = 0$ is equivalent to the unconditional DDIM, $w = 1$ to DDIM_{ef} (no CFG), $w = 2$ to DDIM_{ef}, and $w > 2$ is similar to DDIM_{ef}, but with a larger guidance scale. The two drusen between the lime vertical lines seem to grow with a larger guidance scale.

images and overall coherence, compared to a diffusion-based approach without *en face* information as conditional input. Our method specifically improves the reconstruction of superficial blood vessels and geographic atrophy (see Fig. 7.3 and 7.C.1). Furthermore, our diffusion models demonstrate visually sharper images than tricubic interpolation.

In terms of the classical image similarity metrics MSE, SSIM, and PSNR, which are sub-optimal for evaluating super-resolution methods [308], tricubic interpolation performed roughly the same as our diffusion models. In terms of $\text{LPIPS}_{\text{cor}}$, $\text{LPIPS}_{\text{sag}}$, and $\text{LPIPS}_{\text{efproj}}$, our diffusion models outperformed tricubic interpolation but slightly underperformed in terms of $\text{LPIPS}_{\text{axi}}$. This finding is in line with our visual observations in Appendix Section 7.D.

In terms of all reported quantitative metrics, using CFG to guide the diffusion model more towards the information in the *en face* image either slightly decreased performance, or showed no effect (see Table 7.1). Paradoxically, we found using CFG could lead to struc-

tural features that more closely resembled those in the high-resolution reference than when CFG was not employed. An example of this effect is shown in Fig. 7.5. Besides, this example shows that setting the guidance scale w too high can lead to exaggerated structural features (*e.g.,* too large drusen) and artifacts.

This study has limitations. (1) Our DDIM models can sometimes introduce imaging artifacts (see Appendix Section 7.E). (2) When visually inspecting the dataset, we found that the registration information between OCT B-scans and the SLO image was not always perfect, possibly not allowing the model to utilize the SLO/OCT mapping well sometimes. An improved registration strategy will likely improve generation results and learning speed. (3) Although we present various qualitative results, we only evaluated our approach quantitatively using image similarity metrics. Therefore, we cannot draw conclusions about whether our super-resolution approach improves biomarker quantification. This requires a vast amount of manual annotation labor or a reliable segmentation model for the type of OCT data used in this work. As we did not have access to these resources, we leave this evaluation to future work. (4) Hallucination is a large risk of most generative models, including diffusion models [318]. In the context of medical imaging, generative models like ours risk the generation of non-existent lesions (leading to false positives), inflating them (leading to over-quantification), removing them (leading to false negatives), or shrinking them (leading to under-quantification). Although our *en face* conditioning mechanism may reduce hallucinations by providing more context to make well-informed generative decisions, sufficient empirical evidence of our method completely preventing this is lacking. (5) The sampling time for our diffusion model is relatively long. However, approaches exist to reduce this sampling time [12, 313].

We only explored the effect of conditioning OCT super-resolution diffusion models with near-infrared SLO images. Future work could include more *en face* modalities as conditional information, such as CFP and FAF. As images from those modalities would likely provide additional information than SLO images, we expect this could lead to more accurate super-resolution models. Other metadata, such as functional vision exam data and OCT scans from other devices or protocols, may contain even more useful information.

In our current implementation, we resize and vertically repeat the SLO image patch, enabling concatenation in the channel dimension with the OCT volume. This turns the 2D SLO image into a volume that is processed by 3D convolutions. This is computationally inefficient and the initial resizing can lead to information loss. Future work could focus on designing an architecture that more effectively leverages this multimodal data, possibly using a separate encoder for *en face* images and a cross-attention mechanism to combine the features from the different encoders.

The approach of using relatively high-resolution 2D images from a certain modality to condition diffusion models for super-resolving 3D data from another modality could potentially be applied in other medical domains. For example, using high-resolution 2D X-ray imaging as conditional information for super-resolution CT or MRI scans may be an interesting future direction.

In conclusion, we have shown the feasibility of conditioning super-resolution diffusion models to reduce anisotropy in volumetric images with additional and readily available image data, enabling well-informed generative decisions. Specifically, we showed this in the context of OCT super-resolution conditioning on *en face* images. We think this can be an important next step towards standardized high-quality OCT and other volumetric imaging, leading to more consistent measurements – obtained from either downstream manual quantifications or machine learning models – within and across datasets, studies, and clinical practices. Furthermore, our approach could facilitate the trustworthiness of generative models and their regulatory approval by mitigating the risk of hallucinations compared to uninformed super-resolution models.

Acknowledgements

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116076. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. This communication reflects the author's view and neither IMI nor the European Union and EFPIA are responsible for any use that may be made of the information contained therein.
MACUSTAR consortium members: H. Agostini, I. D. Aires, L. Altay, R. Atia, F. Bandello, P. G. Basile, J. Batuca, C. Behning, M. Belmouhand, M. Berger, A. Binns, C. J. F. Boon, M. Böttger, J. E. Brazier, C. Carapezzi, J. Carlton, A. Carneiro, A. Charil, R. Coimbra, D. Cosette, M. Cozzi, D. P. Crabb, J. Cunha-Vaz, C. Dahlke, H. Dunbar, R. P. Finger, E. Fletcher, M. Gutfleisch, F. Hartgers, B. Higgins, J. Hildebrandt, E. Höck, R. Hogg, F. G. Holz, C. B. Hoyng, A. Kilani, J. Krätzschmar, L. Kühlewein, M. Larsen, S. Leal, Y. T. E. Lechanteur, D. Lu, U. F. O. Luhmann, A. Lüning, N. Manivannan, I. Marques, C. Martinho, A. Miliu, K. P. Moll, Z. Mulyukov, M. Paques, B. Parodi, M. Parravano, S. Penas, T. Peters, T. Peto, S. Priglinger, R. Ramamirtham, R. Ribeiro, D. Rowen, G. S. Rubin, J. Sahel, C. Sánchez, O. Sander, M. Saßmannshausen, M. Schmid, S. Schmitz-Valckenberg, J. Siedlecki, R. Silva, E. Souied, G. Staurenghi, J. Tavares, D. J. Taylor, J. H. Terheyden, A. Tufail, P. Valmaggia, M. Varano, A. Wolf, N. Zakaria

Chapter appendix

7.A Additional dataset and pre-processing details

For the training and validation set combined, the number of B-scans per OCT volume varied, with 1 volume containing 237 B-scans, 927 volumes containing 241 B-scans, and 29 volumes containing 512 B-scans. The B-scan sizes ranged from 512 $\times$ 496 pixels to 1536 $\times$ 496 pixels, and the physical OCT volume size ranged from 2.9 $\times$ 1.9 $\times$ 2.9 mm^3 to 9.2 $\times$ 1.9 $\times$ 7.7 mm^3. For the test set, all OCT volumes contained 241 B-scans. The B-scan sizes ranged from 512 $\times$ 496 pixels to 1536 $\times$ 496 pixels, and all OCT volumes had a physical size of approximately 8.8 $\times$ 1.9 $\times$ 7.3 mm^3.

Next to the previously described set of OCT volumes, we used the near-infrared confocal SLO images, which Heidelberg Spectralis devices acquire alongside the OCT, as *en face* modality for our conditional diffusion models. The SLO image was registered to the OCT volume according to the physical linkage information between these two images that was provided by the camera software. We subsequently cropped and resized these SLO images to the same width and height as, respectively, the width and depth from their corresponding OCT volumes.

For some OCT volumes in the dataset, we observed that incidentally adjacent B-scans were not correctly aligned vertically. Therefore, during sampling, B-scans were registered vertically using a grid search for the vertical translation amount and MSE as a cost function, based on a flattened representation of the B-scans (collapsed into columns by averaging over the x-axis).

7.B Overlapping patches and inpainting

During sampling of full volumes, we use overlapping patches, facilitated through inpainting with RePaint [310]. Inpainting is the task of filling in new content in a specific part of an image, which can be defined by a binary mask. We refer to the image part that needs to be filled in as "unknown" and the other part as "known". During each timestep t in the sampling process, RePaint combines the "known" image part from the input image, which has been noised to the appropriate noise level of timestep $t-1$, with the "unknown" part from the denoised image $\mathbf{x}_{t-1}$ (see the bottom part of Fig. 7.B.1). When we generate a com-

plete volume using this patch overlapping strategy, the "known" region is defined as the area previously generated from an adjacent patch, combined with slices from the original low-resolution image volume (see the top part of Fig. 7.B.1). Besides, the patch size during sampling is larger than the one used during training, as we empirically found this to improve the coherence and fidelity of the generated volumes.

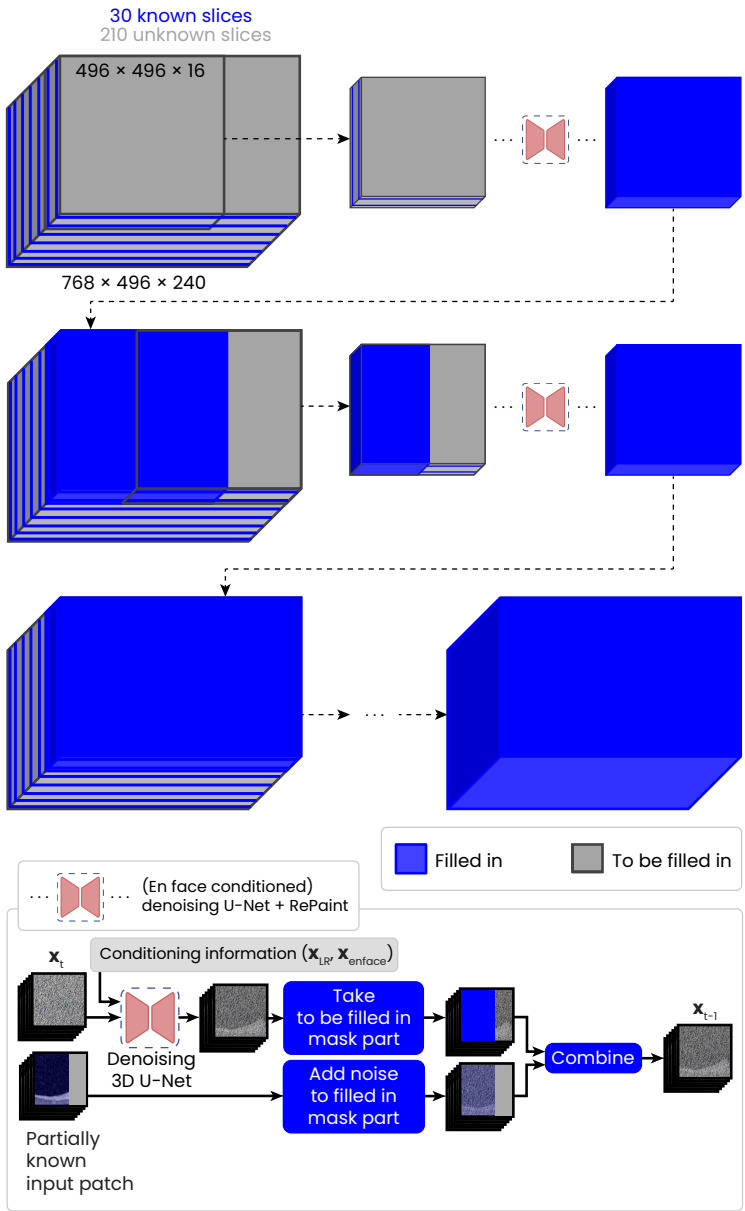


Figure 7.B.1: Overview of our overlapping strategy facilitated through RePaint [310].

7.C Additional results

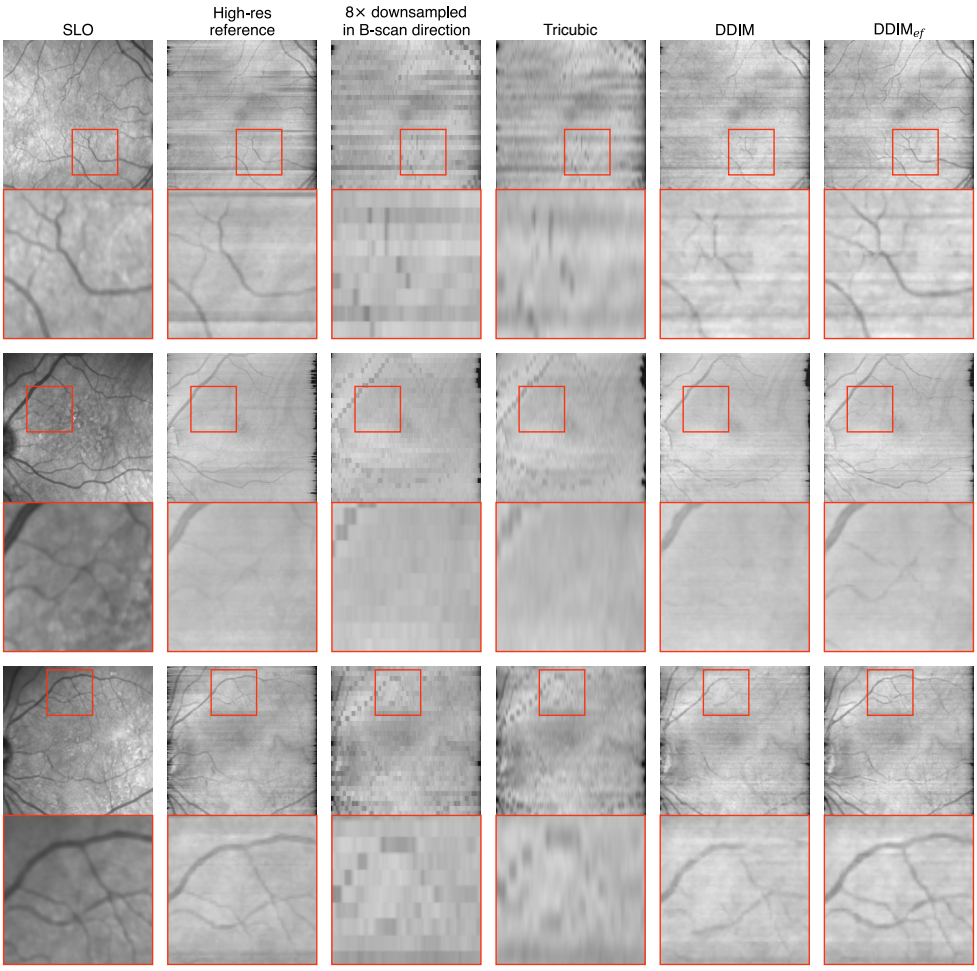


Figure 7.C.1: Additional examples of *en face* projections of the OCT volumes generated using tricubic interpolation, the unconditional diffusion model DDIM, and our proposed *en face* conditioned diffusion model DDIM$_{\text{ef}}$ (presentation similar to Fig. 7.3). These three projections are shown in the last three columns and were all generated from the 8 × downsampled (in the B-scan direction) volume as input, which is shown in the third column. The first and second rows show the corresponding high-resolution (high-res) reference and scanning laser ophthalmoscopy (SLO) image, respectively. A separate example from a different test set patient is shown in each row. The top images in each row show the full image. Zoomed-in versions of the image crops (red boxes) are shown at the bottom of each row.

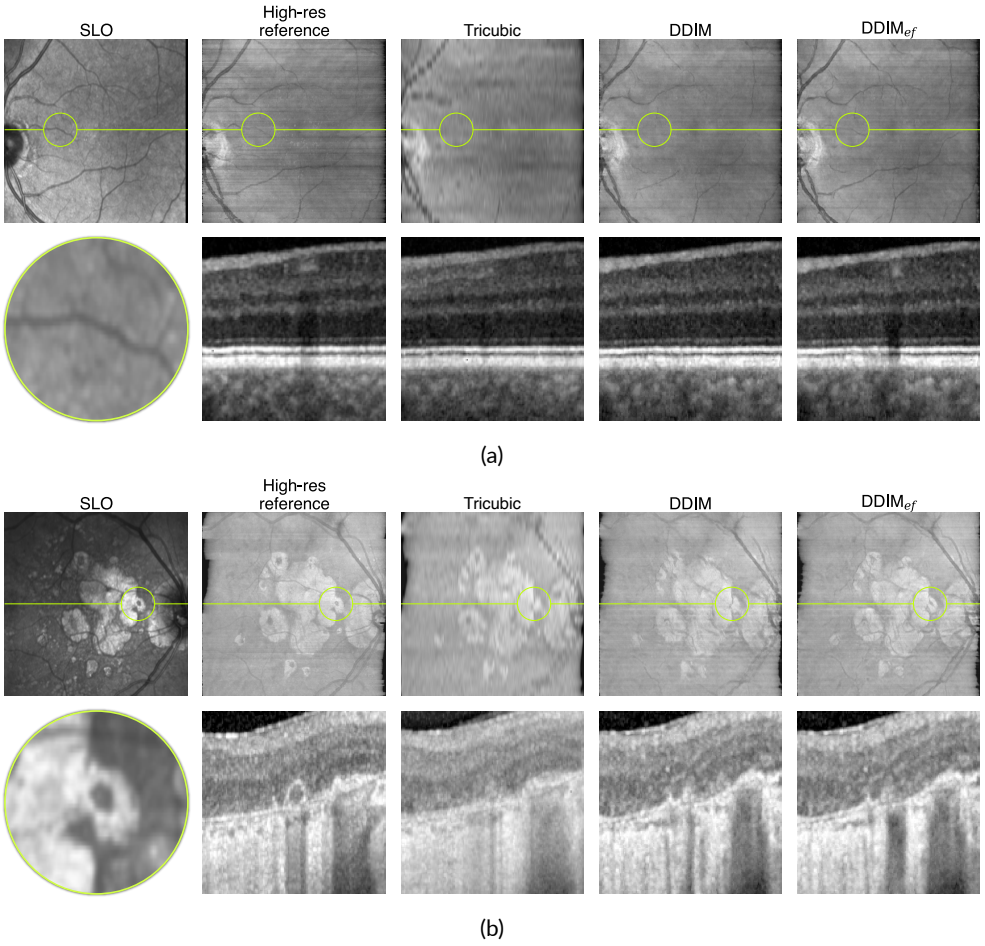


Figure 7.C.2: Additional examples showing the effect of using DDIM$_{ef}$, compared to leaving out the *en face* information (DDIM), and tricubic interpolation (presentation similar to Fig. 7.4). In the top rows, the scanning laser ophthalmoscopy (SLO) image is shown on the left, followed by the *en face* projections of each OCT volume. In the bottom row, zoomed-in crops are shown. The crop locations are indicated in the top row with a lime circle. The lime horizontal line corresponds to the B-scan location from the shown crops. **(a)** A blood vessel is only reconstructed by DDIM$_{ef}$. This vessel is also visible in the SLO image. **(b)** The hypertransmission pattern seems to be best reconstructed by DDIM$_{ef}$. The *en face* location of the hypertransmission area is also visible in the SLO image.

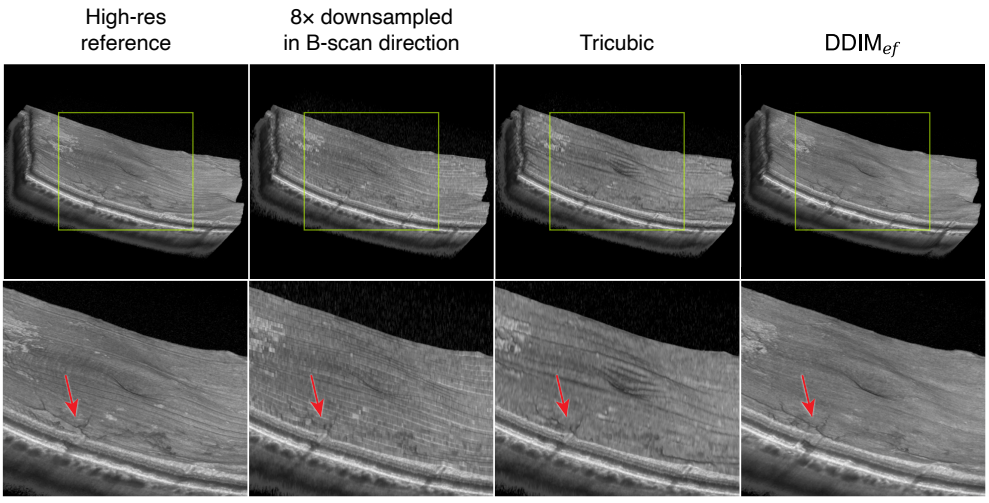


Figure 7.C.3: 3D renders of full OCT volumes, depicted for the high-resolution (high-res) reference, downsampled volume, tricubic interpolated volume, and our proposed method DDIM$_{ef}$. The bottom row shows zoomed-in versions of the renders in the top row. The lime squares in the top row indicate the zoomed-in area. The red arrows point to a vessel on the inner retina.

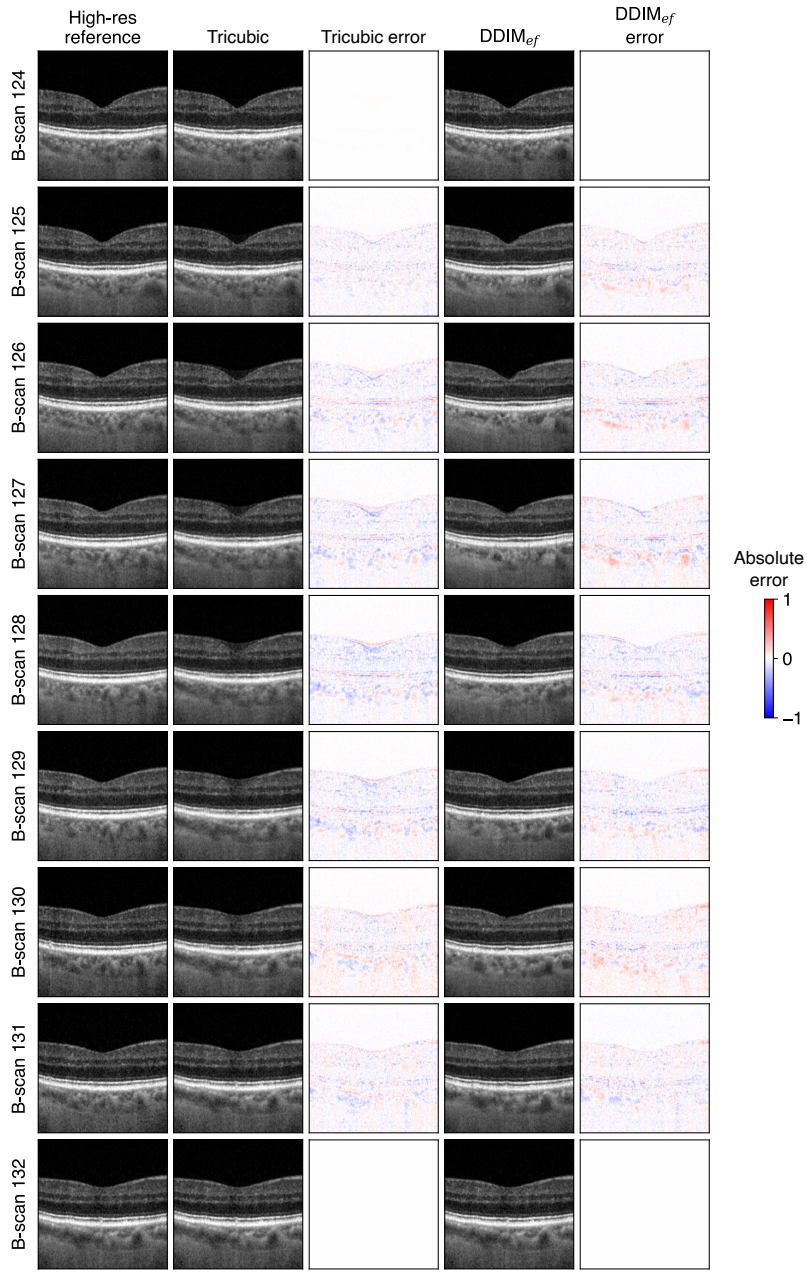


Figure 7.C.4: An example showing a sequence of B-scans, illustrating that the volume generated with tricubic interpolation is more smoothed than the volume generated by our diffusion model DDIM$_{ef}$, which is much sharper. This is most evident around slice 128. The first and last shown slices were already present in the low-resolution OCT volume. The third and fifth columns indicate the absolute error.

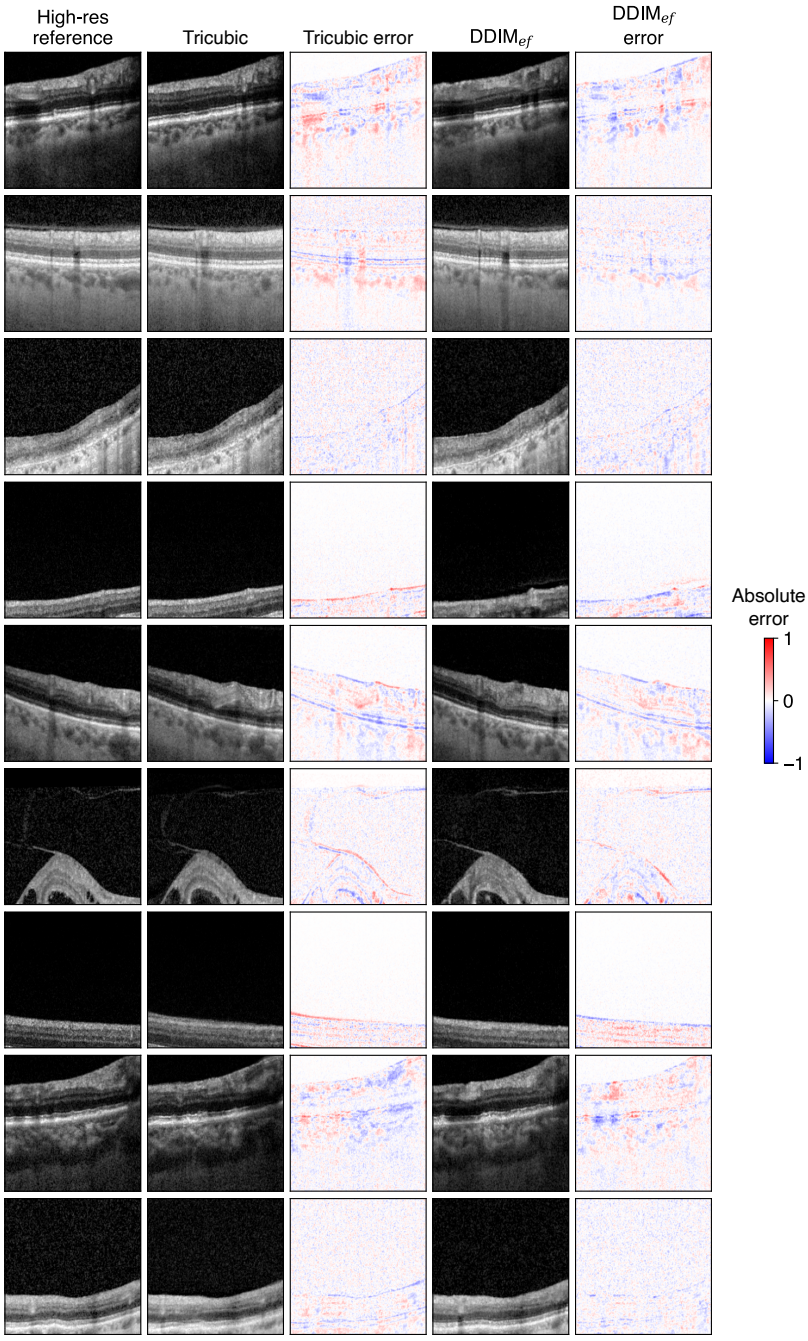


Figure 7.C.5: Randomly picked 256×256 B-scan crops from "unknown" slices in the test set, depicting the origins of absolute errors for tricubic interpolation and DDIM$_{ef}$. The third and fifth columns indicate the absolute error.

7.D LPIPS for different anatomical planes

We found that our diffusion models outperformed interpolation in terms of $\text{LPIPS}_{\text{cor}}$, $\text{LPIPS}_{\text{sag}}$, and $\text{LPIPS}_{\text{efproj}}$, but not in terms of $\text{LPIPS}_{\text{axi}}$. This finding is in line with the visual observations we make when manually inspecting slices from these anatomical planes. This is illustrated in Fig. 7.D.1, in which DDIM_{ef} shows higher visual similarity with the high-resolution reference than the image upsampled with tricubic interpolation for the coronal plane, sagittal plane, and the *en face* projection. For the axial plane, however, we think this is less evident. The LPIPS values, which are also shown in Fig. 7.D.1 for the shown slices, correlate well with these visual observations.

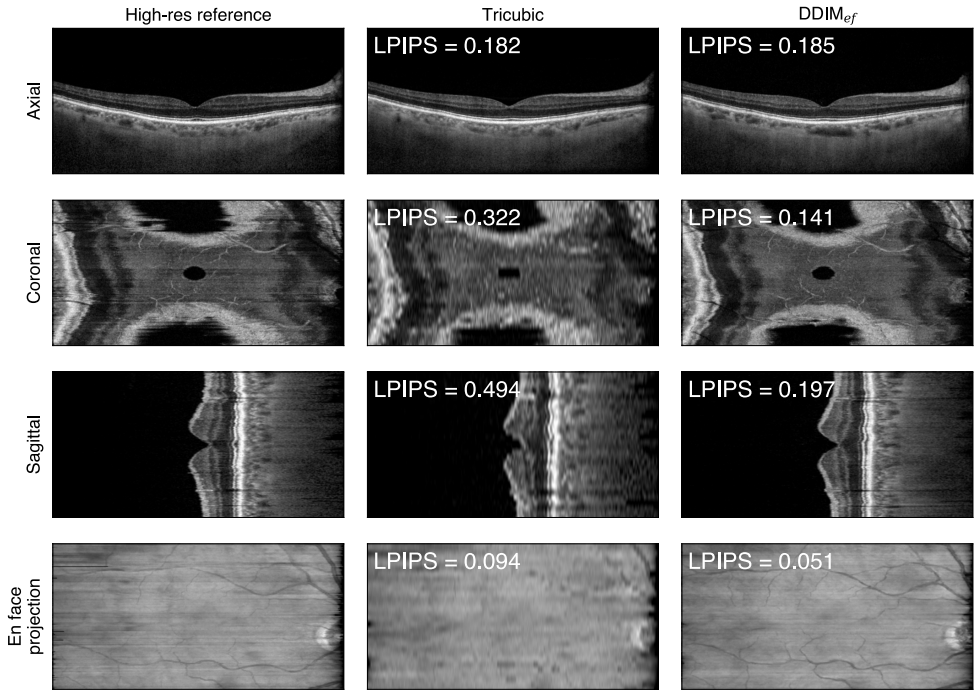
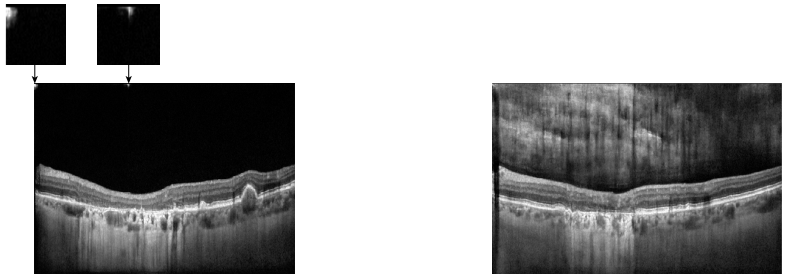


Figure 7.D.1: Examples of slices from the three orthogonal anatomical planes and the *en face* projection image. All shown images originate from the same OCT volume. For tricubic interpolation and DDIM_{ef}, Learned Perceptual Image Patch Similarity (LPIPS) values, computed only using the depicted slices, are shown in the top left corner of each image. For the anatomical planes, the middle slices are shown. For the axial (B-scan) slice, this corresponds to a slice exactly in the middle of two slices that were also present in the low-resolution volume (*i.e.*, the slice was as far away from a "known" slice as possible, which is generally a slice that is more difficult to generate accurately than a slice that is closer to a "known" slice).

7.E Generated imaging artifacts

We found that a number of imaging artifacts can occur when generating volumes using our DDIM approach (see 7.E.1). The occurrence frequency depends on the artifact type. For example, small groups of white pixels are sometimes generated near the top of B-scans (see an example in Fig. 7.E.1a). If they are present, they always seem to be located in the top left of the sampling patches. Furthermore, sometimes very large, bright areas are generated in the vitreous body (see an example in Fig. 7.E.1b). This mainly seems to occur when two adjacent B-scans in the low-resolution volume were not registered well.



(a) (b)

Figure 7.E.1: Artifact types that sometimes occur in volumes generated by our diffusion models. These artifacts were neither present in the low-resolution input nor the high-resolution reference. **(a)** Small groups of white pixels in the top left of sampling patches. **(b)** White areas in the vitreous body. This artifact type mainly seems to occur when two adjacent B-scans in the low-resolution volume were not registered well.

7.F Ethical approval

Ethical approval of the clinical study data has been obtained, as described in previous works describing the MACUSTAR study [319].

8

General discussion

This thesis examines the robustness of deep learning (DL) models for medical image analysis within radiology and ophthalmology applications. We investigated model performance under varying degrees of data shift in several clinically relevant settings. In addition, we evaluated methods for identifying out-of-distribution (OOD) samples and explored approaches to mitigate performance degradation resulting from data shifts. In this chapter, we provide a general discussion of the main results and contributions of this thesis and outline future research directions. We summarize our key findings on model robustness, discuss the clinical implications of our studies, and touch upon our efforts to make the models in this thesis available to other researchers.

8.1 Improving robustness

As demonstrated by previous research, DL models can exhibit high performance in controlled environments but are susceptible to failure under distribution shifts [23, 208]. The models in Chapters 2, 3, 4, 5, and 6 similarly showed strong performance for classification and segmentation tasks on test data similar to the training data.

In Chapter 2 and Chapter 3, we observed good generalization of our COVID-19 models could be achieved when applied to external data from, respectively, a teaching hospital and the publicly available iCTCF dataset [110]. In some respects, the external datasets were different compared to the training data, such as the patient population and the reference standard. However, we found these discrepancies to be relatively limited in terms of the overall computed tomography (CT) scan appearances, especially compared to the discrepancies found between optical coherence tomography (OCT) datasets, which we will discuss later in this section. Possible reasons for this include the more standardized nature of the quantitative Hounsfield scale in CT and the general use of the standardized data format Digital Imaging and Communications in Medicine (DICOM) in radiology. Consequently, we considered these external datasets to at least closely align with our training distribution.

In Chapter 4, we also found the algorithms from the teams participating in the Artificial Intelligence for Robust Glaucoma Screening (AIROGS) challenge to generalize well to both internal and external test sets. The algorithms performed similarly for referable glaucoma classification to 20 expert ophthalmologists and optometrists on the AIROGS test set. They generally achieved even higher performance on the external datasets. This could be attributed to the large and diverse training set, which included color fundus photographs (CFPs) from many different sites and devices. Furthermore, the external test sets likely contained fewer borderline cases and more patients with advanced glaucoma stages, generally allowing for less ambiguous interpretations.

Similarly, the DL models for OCT applications, specifically the age-related macular degeneration (AMD) classification model in Chapter 5, and the underlying model from Liefers *et al.* for the segmentation of AMD biomarkers [68] used in Chapter 6, demonstrated strong performance on internal test sets. However, we observed substantial performance degradations when these models were applied to external datasets from diverse OCT devices.

8.1.1 Robustness through uncertainty

Modern DL models are prone to produce inaccurate and over-confident probability outputs [28, 320, 321]. To address this, uncertainty estimation has emerged and grown as a research field over the last few years, aiming to answer the question: *Can our models communicate to the outside world what they know and do not know?* Furthermore, uncertainty estimation techniques are known to be effective at OOD detection [215]. Given these advancements, we explored various ways of enhancing the robustness of DL models through uncertainty.

When organizing the AIROGS challenge (Chapter 4), we only provided gradable CFPs for model training. However, the unfiltered, real-world test set included both gradable and ungradable data. The participating teams were tasked with identifying this ungradable data, encouraging the development of general methods that are robust to any ungradable or unexpected input data. The participating teams proposed diverse solutions, predominantly based on uncertainty estimation or OOD detection methods. These ranged from techniques based on the image reconstruction error of (variational) auto-encoders [172,183] to methods using multiple model forwards passes, such as Monte-Carlo dropout [193,194], ensembles [192,321], or test-time augmentation [187,247]. Notably, the top three teams all implemented strategies based on domain knowledge. They developed DL models to detect or segment relevant anatomical features in the images, such as the optic disc or retinal vessels. The confidence or uncertainty of these models then served as a gradability score. This finding raises questions about the effectiveness of domain knowledge-based OOD detection methods for a wider range of medical imaging applications, a possibly interesting direction for future research.

In Chapter 5, we introduced another approach leveraging uncertainty to improve DL robustness: Uncertainty-Based Instance eXclusion (UBIX). UBIX addresses the robustness of classification models to data with localized OOD characteristics, such as local imaging artifacts. In OCT datasets, device-specific artifacts were found to significantly degrade performance. By suppressing the influence of these artifacts on volume-level predictions, UBIX can mitigate this degradation without requiring training on data with such artifacts. UBIX was particularly effective when using deep ensembles as its underlying uncertainty esti-

mation method. This is in line with previous research that identified deep ensembles as an overall effective technique when compared to other predictive uncertainty quantification methods under data shift [208]. Nevertheless, UBIX is expected to be compatible with most uncertainty estimation techniques and may benefit from the integration with even more effective OOD detection approaches. We validated UBIX for AMD staging in OCT, a task well-suited for UBIX due to the variability and lack of standardization in OCT imaging protocols and devices. Potential applications extend to other medical imaging domains where dataset-specific artifacts can occur.

8.1.2 Generalization through generation

Recent advancements in generative DL have shown substantial progress, with models that excel at a wide variety of tasks, including text [11], image [312], video [314], and audio [322] generation. Generative models can be conditioned on specific information, such as text or images [12, 293, 323], allowing generations guided by additional information. In this thesis, we aimed to enhance model robustness through conditional generative models that realign shifted distributions of OCT datasets to a desired distribution.

In Chapter 6, we trained a Cycle-Consistent Generative Adversarial Network (CycleGAN) to adapt the characteristics of OCT scans from one device to those in the training data, aiming to enhance the robustness of a downstream segmentation model for delineating incomplete retinal pigment epithelium (RPE) and outer retinal atrophy (iRORA) and complete RPE and outer retinal atrophy (cRORA). The downstream segmentation model was trained on data from a Topcon OCT device whereas we applied it to OCT scans from both Zeiss Cirrus and Heidelberg Spectralis devices. In particular, the contrast and signal-to-noise ratio in scans from the latter device differed because of the use of B-scan averaging. Our approach demonstrated good generalization and superior consistency across devices compared to human annotations. We observed a 16% improvement in quadratic weighted Cohen's κ score when introducing the CycleGAN model. However, establishing statistically significant differences in this dataset was challenging due to low prevalence of iRORA and cRORA. Future research should focus on the evaluation with more iRORA and cRORA cases to enhance the statistical power and reliability of the findings.

In Chapter 7, we leveraged recent advancements in diffusion models as a generative framework. Previous research has demonstrated that diffusion models are highly effective at utilizing information from other modalities, such as text, to generate high-fidelity images and videos [324]. We extended this concept to standardize 3D OCT scans to high-quality data based on 2D *en face* images by increasing their B-scan density. We

think this standardization of OCT scans is important, as the variability in B-scan spacing across datasets and vendors can result in highly inconsistent quantitative measurements [297–300]. Our work demonstrated promising results in both qualitative and quantitative evaluations. We quantitatively evaluated our approach in terms of classical and perceptual image similarity metrics. However, future research should focus on evaluating the benefits of these approaches for consistently and quantitatively measuring biomarkers in OCT.

8.1.3 Model dimensionality

Throughout this thesis, our DL models were based on convolutional neural networks (CNNs). When working with volumetric images like CT and OCT, we found that the dimensionality of the convolutional kernels in these networks significantly influenced model robustness. In our dimensionality experiments for CT data, detailed in Chapter 3, we observed that 3D architectures consistently outperformed their 2D counterparts. The isotropic nature of the volumetric CT scans utilized in this work is well-suited for vanilla 3D CNNs. It allows the model to use information from the additional dimension, mirroring the workflow of radiologists when they consider multiple orthogonal planes during the manual assessment of a scan.

For OCT scans, however, vanilla 3D CNNs may not always be the optimal design choice, even though these scans are also volumetric. OCT scans are often anisotropic, typically exhibiting larger B-scan spacing compared to the spacing within B-scans. Additionally, the variability in B-scan spacing across different OCT datasets hampers the generalization of trained 3D convolutional kernels, particularly when addressing robustness to data from different devices. In Chapter 5, we found that a 3D CNN baseline performed well on internal test data, but its performance dropped drastically on external data. This performance drop was significantly less pronounced in 2D CNN-based architectures.

To effectively utilize information from multiple B-scans in anisotropic OCT, a hybrid approach has been proposed by de Fauw *et al.* [205], involving 2D convolutions in the xy-plane combined with 1D convolutions in the z-direction in the deeper layers of the network. This strategy was also implemented in the segmentation model from Liefers *et al.* [68] that we used in Chapter 6. However, this method does not fully address all issues related to poor generalization across different OCT devices. De Fauw *et al.* demonstrated this limitation when their proposed model was applied to external data from a different device type than the one used for training [205]. This underscores the need for data standardization techniques, such as the approach proposed in Chapter 7.

8.1.4 Concept shift

A less explored aspect of robustness is concept shift [325], where the robustness modifier (see Section 1.2) is not the deployment data but the definition of the labels. Changes in the definition of a particular disease stage, for instance, can result in a decline in the measured model performance. This phenomenon was observed in this thesis as a limiting factor for the accurate estimation of external performance.

For example, in Chapter 3, CO-RADS labels were not available for the external test set, which was labeled with a different classification system. We harmonized the classification systems by binarizing the five-grade scoring system, which limited us in effectively evaluating the models for the originally multi-class task. We encountered a similar issue in Chapter 5 while standardizing AMD staging scales. In Chapter 4, we used a binary classification system (glaucoma vs. no referable glaucoma) for training, but the definitions in the external evaluation sets, although similar, were not always identical. Notably, the iRORA/cRORA detection model from Chapter 6 was based on more granular structural features in OCT, likely making this approach more robust to future definition changes of iRORA and cRORA. These observations highlight the necessity for well-defined reference standards and their consistent adoption in research. This would not only limit concept shifts, but also enhance the comparability of algorithms across different studies.

8.2 Clinical implications

8.2.1 Glaucoma screening

Population-wide screening for glaucoma is currently considered neither beneficial nor cost-effective [326, 327]. However, targeted screening in high-risk populations, including individuals with a family history of glaucoma, elderly individuals, and people with an African American or Hispanic ethnic background, has demonstrated greater efficacy [328]. The implementation of broader screening programs is hindered by the subjectivity and low specificity of many screening methods [328]. We expect that the relatively cheap imaging modality of fundus photography could substantially enhance the widespread implementation of glaucoma screening. However, the subjectivity associated with CFP interpretation has been a constraining factor in its application for glaucoma screening [328]. In the AIROGS challenge (Chapter 4), we used metrics based on high specificities, while limiting subjectivity in our reference standard through the thorough training and continuous monitoring of ophthalmologists and optometrists. Our results demonstrate that artificial intelligence (AI) algorithms can achieve comparable performance to human graders.

Previous AI solutions for CFP-based glaucoma screening suffered from low generalizability and the lack of transparency regarding which discriminatory factors algorithms use to detect glaucoma [328]. We have addressed the generalization issue by releasing the largest public CFP dataset for glaucoma screening, which includes patients from diverse ethnic backgrounds and numerous screening sites, and demonstrated robust performance on external test sets. A recent AIROGS follow-up challenge, the JustRAIGS challenge*, aims to address the explainability issue by focusing on the identification of ten specific glaucoma-related features.

An alternative approach to improve screening accuracy involves the use of more advanced imaging techniques such as OCT. However, cost and accessibility limitations restrict this deployment in a larger scope. As OCT technology continues to evolve and potentially become more affordable, these limitations may be mitigated in the future.

Another promising approach is to integrate automated glaucoma screening algorithms into existing diabetic retinopathy (DR) screening programs, where AI-based color fundus analysis pipelines are already in place. Given the high performance of the AIROGS algorithms, one would intuitively expect that setting a very strict operating point in such a setting – one that avoids virtually any false positive – could still result in a significant number of true positives with no additional costs related to overdiagnosis or the screening operation itself. However, studies investigating the cost-effectiveness of such approaches are currently still lacking [329].

8.2.2 Managing AMD progression

Several applications and extensions of the models for AMD in OCT presented in this thesis are conceivable. Firstly, generalizable models such as the one presented in Chapter 6 for iRORA and cRORA segmentation can be of instrumental value in gaining better insights into the disease progression from intermediate AMD to late AMD.

Utilizing these models in large-scale, longitudinal studies such as Macustar [266], a multi-center study on intermediate AMD, may facilitate the development of novel interventions that could inhibit the progression of the disease. Such models may be beneficial for conducting structure-function analyses, helping to elucidate which structural changes are associated with vision loss. Additionally, implementing these segmentation models and using their output as endpoints in clinical trials could improve the quantitative assessment of intervention efficacy, resulting in more reliable trial outcomes.

The integration of AI models to assess the progression of wet AMD could revolutionize the management and clinical approval of anti-vascular endothelial growth factor (anti-

*https://justraigs.grand-challenge.org/

VEGF) injections. Currently, wet AMD treatment strategies are predominantly based on pro re nata or treat and extend regimens. We think that complementing these strategies with AI models – based on quantifiable growth measurements – may significantly enhance the efficacy of these therapies, both during clinical trials and in clinical practice. As treatments for geographic atrophy are becoming available [64], such strategies could also be applied in the treatment of dry AMD patients.

8.3 Democratization and reproducibility through open source science

There is a trend in machine learning towards larger models trained on increasingly bigger datasets, seemingly outpacing the reduction in computational costs. This makes it difficult for smaller organizations to develop such models, concentrating the ownership of advanced AI models within large commercial entities. Often, these models will not be made publicly available and open source, hampering academic researchers in using these models for less commercially viable, yet societally valuable, applications. Furthermore, the lack of transparency and scientific documentation can introduce unknown biases and risks that negatively impact society. Therefore, it is crucial for scientific AI research to maintain an open source and scientific character. This is especially important for foundation models, which pose a promising next step for robust and generalizable DL models [14, 330].

In this thesis, we have made our code and models publicly accessible whenever feasible. We have made our algorithms easily reproducible through the Grand Challenge platform (grand-challenge.org). Through a web interface, users can freely apply their own data to our CORADS-AI model described in Chapter 2 and 3, as well as the participating algorithms from the AIROGS challenge in Chapter 4, and the UBIX model discussed in Chapter 5. For the remaining algorithms in this thesis, we are still in the process of making them compatible with the Grand Challenge platform. Additionally, the evaluation code for the AIROGS challenge is publicly available and the automated evaluation of new submissions is still supported through the Grand Challenge platform. Relevant links to code and algorithms are provided in the respective chapters of this thesis.

8.4 Concluding remarks

We have presented multiple approaches to enhance the robustness of DL models in ophthalmology and radiology settings. Specifically, we developed models for the detection and quantification of COVID-19, the classification of glaucoma, and the grading and quantifica-

tion of AMD. These models demonstrated good generalization to data distributions that were similar to the distribution of the training data. In scenarios where the data significantly diverged from the training distribution, we identified methods to accurately detect these occurrences and strategies to mitigate the resulting performance degradation.

Incorporating mechanisms to increase robustness is important for the transition of DL models from research settings to real-world medical applications. However, additional challenges need to be addressed to ensure the clinical utility and trustworthiness of these models. These challenges include the integration of models into clinical workflows, which can be particularly challenging within the closed systems of ophthalmology, cost-effectiveness analyses, and prospective validation studies to assess the clinical impact of these systems.

Bibliography

[1] I. R. Schwab, "The evolution of eyes: major steps. the Keeler lecture 2017: centenary of Keeler Ltd," *Eye*, vol. 32, no. 2, pp. 302–313, 2018.

[2] C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[3] S. A. Papert, "The summer vision project," 1966.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv:14094842v1*, 2014.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015, pp. 1026–1034.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 9351, 2015, pp. 234–241.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[13] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[14] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[15] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, and S. Chintala, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024. [Online]. Available: https://pytorch.org/assets/pytorch2-2.pdf

[16] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow, Large-scale machine learning on heterogeneous systems," Nov. 2015.

[17] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/google/jax

[18] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 12 2017. [Online]. Available: https://arxiv.org/abs/1702.05747

[19] P. M. Cheng, E. Montagnon, R. Yamashita, I. Pan, A. Cadrin-Chênevert, F. Perdigón Romero, G. Chartrand, S. Kadoury, and A. Tang, "Deep learning: an update for radiologists," *Radiographics*, vol. 41, no. 5, pp. 1427–1445, 2021.

[20] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.

[21] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. Van Boven, R. Vink *et al.*, "Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge," *Nature medicine*, vol. 28, no. 1, pp. 154–163, 2022.

[22] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij, "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence," *European radiology*, vol. 31, pp. 3797–3804, 2021.

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

[24] T. Freiesleben and T. Grote, "Beyond generalization: a theory of robustness in machine learning," *Synthese*, vol. 202, no. 4, p. 109, 2023.

[25] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[26] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2065–2081, 2020.

[27] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[28] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[29] M. Prokop, W. van Everdingen, T. van Rees Vellinga, J. Quarles van Ufford, L. Stoger, L. Beenen, B. Geurts, H. Gietema, J. Krdzalic, C. Schaefer-Prokop, B. van Ginneken, M. Brink, and COVID-19 Standardized Reporting Working Group of the Dutch Radiological Society, "CO-RADS: a categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation," *Radiology*, vol. 296, no. 2, pp. E97–E104, 4 2020.

[30] W. Yang, A. Sirajuddin, X. Zhang, G. Liu, Z. Teng, S. Zhao, and M. Lu, "The role of imaging in 2019 novel coronavirus pneumonia (COVID-19)," *European Radiology*, vol. 30, no. 9, pp. 1–9, 2020.

[31] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:2003.13865*, 2020.

[32] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, no. 2, pp. 65–72, 2020.

[33] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu *et al.*, "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis," *European Respiratory Journal*, vol. 56, no. 2, 2020.

[34] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Zha *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, early access.

[35] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," *European radiology*, vol. 31, no. 8, pp. 1–9, 2021.

[36] B. Wang, S. Jin, Q. Yan, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng, W. Sun, L. Lan, W. Zhang, X. Mu, C. Shi, Z. Wang, J. Lee, Z. Jin, M. Lin, H. Jin, L. Zhang, J. Guo, B. Zhao, Z. Ren, S. Wang, W. Xu, X. Wang, J. Wang, Z. You, and J. Dong, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system," *Applied Soft Computing*, vol. 98, p. 106897, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494620308358

[37] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song *et al.*, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, 2020.

[38] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, and D. Qian, "Prior-attention residual learning for more discriminative COVID-19 screening in CT images," *IEEE Transactions on Medical Imaging*, 2020.

[39] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[40] N. Lessmann, C. I. Sanchez, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke, B. Geurts, H. A. Gietema, M. Groeneveld, L. van Harten, N. Hendrix, W. Hendrix, H. J. Huisman, I. Isgum, C. Jacobs, R. Kluge, M. Kok, J. Krdzalic, B. Lassen-Schmidt, K. van Leeuwen, J. Meakin, M. Overkamp, T. van Rees Vellinga, E. M. van Rikxoort, R. Samperna, C. Schaefer-Prokop, S. Schalekamp, E. T. Scholten, C. Sital, L. Stöger, J. Teuwen, K. Vaidhya Venkadesh, C. de Vente, M. Vermaat, W. Xie, B. de Wilde, M. Prokop, and B. van Ginneken, "Automated assessment of CO-RADS and chest CT severity scores in patients with suspected COVID-19 using artificial intelligence," *Radiology*, 2020.

[41] W. Ning, S. Lei, J. Yang, Y. Cao, P. Jiang, Q. Yang, J. Zhang, X. Wang, F. Chen, Z. Geng *et al.*, "iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia," *Research Square*, 2020.

[42] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung *et al.*, "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[43] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su *et al.*, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.

[44] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan, "Coronavirus detection and analysis on chest CT with deep learning," *arXiv preprint arXiv:2004.02640*, 2020.

[45] X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi, "COVID-AL: The diagnosis of COVID-19 with deep active learning," *Medical Image Analysis*, vol. 68, p. 101913, 2020.

[46] S. A. Harmon, T. H. Sanford, S. Xu, E. B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou *et al.*, "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nature communications*, vol. 11, no. 1, pp. 1–7, 2020.

[47] A. C. van Trigt, "De speculo oculi ejusque usu," Ph.D. dissertation, Universiteit Utrecht, 1953.

[48] F. Muirhead, "Adrien van trigt, and the first published ophthalmoscopic images (de speculo oculi, 1853)," *Journal of Medical Biography*, vol. 29, no. 4, pp. 221–228, 2021.

[49] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[50] C. R. Keeler, "The ophthalmoscope in the lifetime of Hermann von Helmholtz," *Archives of ophthalmology*, vol. 120, no. 2, pp. 194–201, 2002.

[51] R. H. Webb and G. W. Hughes, "Scanning laser ophthalmoscope," *IEEE Transactions on Biomedical Engineering*, no. 7, pp. 488–492, 1981.

[52] R. H. Webb, G. W. Hughes, and F. C. Delori, "Confocal scanning laser ophthalmoscope," *Applied optics*, vol. 26, no. 8, pp. 1492–1499, 1987.

[53] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[54] P. J. Rosenfeld, "Optical coherence tomography and the development of antiangiogenic therapies in neovascular age-related macular degeneration," *Investigative ophthalmology & visual science*, vol. 57, no. 9, pp. OCT14–OCT26, 2016.

[55] S. Aumann, S. Donner, J. Fischer, and F. Müller, "Optical coherence tomography (OCT): principle and technical realization," *High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics*, pp. 59–85, 2019.

[56] E. A. Swanson and J. G. Fujimoto, "The ecosystem that powered the translation of OCT from fundamental research to clinical and commercial impact," *Biomedical Optics Express*, vol. 8, no. 3, pp. 1638–1664, 2017.

[57] M. J. Burton, J. Ramke, A. P. Marques, R. R. Bourne, N. Congdon, I. Jones, B. A. A. Tong, S. Arunga, D. Bachani, C. Bascaran *et al.*, "The lancet global health commission on global eye health: vision beyond 2020," *The Lancet Global Health*, vol. 9, no. 4, pp. e489–e551, 2021.

[58] R. Bourne, J. D. Steinmetz, S. Flaxman, P. S. Briant, H. R. Taylor, S. Resnikoff, R. J. Casson, A. Abdoli, E. Abu-Gharbieh, A. Afshin *et al.*, "Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study," *The Lancet Global Health*, vol. 9, no. 2, pp. e130–e143, 2021.

[59] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[60] J. B. Jonas, C. M. G. Cheung, and S. Panda-Jonas, "Updates on the epidemiology of age-related macular degeneration," *Asia-Pacific Journal of Ophthalmology*, vol. 6, no. 6, pp. 493–497, 2017.

[61] F. L. Ferris III, C. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, S. R. Sadda, B. I. for Macular Research Classification Committee *et al.*, "Clinical classification of age-related macular degeneration," *Ophthalmology*, vol. 120, no. 4, pp. 844–851, 2013.

[62] R. H. Guymer, P. J. Rosenfeld, C. A. Curcio, F. G. Holz, G. Staurenghi, K. B. Freund, S. Schmitz-Valckenberg, J. Sparrow, R. F. Spaide, A. Tufail *et al.*, "Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: classification of atrophy meeting report 4," *Ophthalmology*, vol. 127, no. 3, pp. 394–409, 2020.

[63] Y. Barak, W. J. Heroman, and T. H. Tezel, "The past, present, and future of exudative age-related macular degeneration treatment," *Middle East African journal of ophthalmology*, vol. 19, no. 1, pp. 43–51, 2012.

[64] K. G. Csaky, J. M. Miller, D. F. Martin, and M. W. Johnson, "Drug approval for the treatment of geographic atrophy: how we got here and where we need to go," *American Journal of Ophthalmology*, vol. 263, pp. 231–239, 2024.

[65] G. A. Saleh, N. M. Batouty, S. Haggag, A. Elnakib, F. Khalifa, F. Taher, M. A. Mohamed, R. Farag, H. Sandhu, A. Sewelam *et al.*, "The role of medical image modalities and AI in the early detection, diagnosis and grading of retinal diseases: a survey," *Bioengineering*, vol. 9, no. 8, p. 366, 2022.

[66] H. Bogunovic, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, K. Gopinath, A. K. Gostar, K. Jeon, Z. Ji, S. H. Kang, D. D. Koozekanani, D. Lu, D. Morley, K. K. Parhi, H. S. Park, A. Rashno, M. Sarunic, S. Shaikh, J. Sivaswamy, R. Tennakoon, S. Yadav, S. De Zanet, S. M. Waldstein, B. S. Gerendas, C. Klaver, C. I. Sánchez, and U. Schmidt-Erfurth, "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1858–1874, 8 2019.

[67] B. Liefers, J. M. Colijn, C. González-Gonzalo, T. Verzijden, J. J. Wang, N. Joachim, P. Mitchell, C. B. Hoyng, B. van Ginneken, C. C. Klaver *et al.*, "A deep learning model for segmentation of geographic atrophy to study its long-term natural history," *Ophthalmology*, vol. 127, no. 8, pp. 1086–1096, 2020.

[68] B. Liefers, P. Taylor, A. Alsaedi, C. Bailey, K. Balaskas, N. Dhingra, C. A. Egan, F. G. Rodrigues, C. G. Gonzalo, T. F. Heeren *et al.*, "Quantification of key retinal features in early and late age-related macular degeneration using deep learning," *American Journal of Ophthalmology*, vol. 226, pp. 1–12, 2021.

[69] U. Schmidt-Erfurth, G. S. Reiter, S. Riedl, P. Seeböck, W.-D. Vogl, B. A. Blodi, A. Domalpally, A. Fawzi, Y. Jia, D. Sarraf *et al.*, "AI-based monitoring of retinal fluid in disease activity and under therapy," *Progress in retinal and eye research*, vol. 86, p. 100972, 2022.

[70] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[71] P. J. Ernest, M. J. Busch, C. A. Webers, H. J. Beckers, F. Hendrikse, M. H. Prins, and J. S. Schouten, "Prevalence of end-of-life visual impairment in patients followed for glaucoma," *Acta Ophthalmologica*, vol. 91, no. 8, pp. 738–743, 2013.

[72] D. Peters, B. Bengtsson, and A. Heijl, "Lifetime risk of blindness in open-angle glaucoma," *American Journal of Ophthalmology*, vol. 156, no. 4, pp. 724–730, 2013.

[73] A. K. Schuster, C. Erb, E. M. Hoffmann, T. Dietlein, and N. Pfeiffer, "The diagnosis and treatment of glaucoma," *Deutsches Ärzteblatt International*, vol. 117, no. 13, p. 225, 2020.

[74] H. G. Lemij, C. de Vente, C. Sánchez, J. Cuadros, N. Jaccard, and K. Vermeer, "Glaucomatous features in fundus photographs of eyes with 'referable glaucoma' of a large population based labeled data set for training an artificial intelligence (AI) algorithm for glaucoma screening," in *Association for Research in Vision and Ophthalmology*, vol. 63, no. 7. The Association for Research in Vision and Ophthalmology, 2022, pp. 2041–A0482.

[75] S. Simpson, F. U. Kay, S. Abbara, S. Bhalla, J. H. Chung, M. Chung, T. S. Henry, J. P. Kanne, S. Kligerman, J. P. Ko *et al.*, "Radiological society of north america expert consensus document on reporting chest CT findings related to COVID-19: endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200152, 2020.

[76] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezanezhad, "Coronavirus disease 2019 (COVID-19) imaging reporting and data system (COVID-RADS) and common lexicon: a proposal based on the imaging data of 37 studies," *European Radiology*, vol. 30, pp. 4930–4942, 2020.

[77] K. Li, J. Wu, F. Wu, D. Guo, L. Chen, Z. Fang, and C. Li, "The clinical and chest CT features associated with severe and critical COVID-19 pneumonia," *Investigative Radiology*, vol. 55, no. 6, pp. 327–331, 2020.

[78] Y.-C. Chang, C.-J. Yu, S.-C. Chang, J. R. Galvin, H.-M. Liu, C.-H. Hsiao, P.-H. Kuo, K.-Y. Chen, T. J. Franks, K.-M. Huang *et al.*, "Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT," *Radiology*, vol. 236, no. 3, pp. 1067–1075, 2005.

[79] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Appl Intell*, 2020.

[80] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[81] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks," *MedRxiv*, pp. 2020–03, 2020.

[82] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D.-C. Wang, L.-B. Shi *et al.*, "Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT," *Radiology*, vol. 296, no. 3, pp. E156–E165, 2020.

[83] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, and Y. Shi, "Lung infection quantification of COVID-19 in CT images with deep learning," *arXiv preprint arXiv:2003.04655*, 2020.

[84] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang *et al.*, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[85] S. Kundu, H. Elhalawani, J. W. Gichoya, and C. E. Kahn Jr, "How might AI and chest imaging help unravel COVID-19's mysteries?" p. e200053, 2020.

[86] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2664–2675, 2020. [Online]. Available: https://arxiv.org/abs/2004.07443

[87] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo, "Genetic epidemiology of COPD (COPDGene) study design," *COPD*, vol. 7, pp. 32–43, 2010.

[88] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019.

[89] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[90] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." *Nature medicine*, vol. 25, pp. 954–961, 2019.

[91] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[92] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 1–42, 2014.

[93] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The AVA-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020.

[94] H. A. Gietema, N. L. Müller, P. V. N. Fauerbach, S. Sharma, L. D. Edwards, P. G. Camp, H. O. Coxson, and Evaluation of C. O. P. D Longitudinally to Identify Predictive Surrogate Endpoints (E. C. L. I. P. S. E) investigators, "Quantifying the extent of emphysema: factors associated with radiologists' estimations and quantitative indices of emphysema severity using the eclipse cohort," *Academic Radiology*, vol. 18, pp. 661–671, 2011.

[95] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19," *IEEE reviews in biomedical engineering*, vol. 14, pp. 4–15, 2020.

[96] R. Ito, S. Iwano, and S. Naganawa, "A review on the use of artificial intelligence for medical imaging of the lungs of patients with coronavirus disease 2019," *Diagnostic and Interventional Radiology*, vol. 26, no. 5, p. 443, 2020.

[97] Y. Xu, L. Ma, F. Yang, Y. Chen, K. Ma, J. Yang, X. Yang, Y. Chen, C. Shu, Z. Fan *et al.*, "A collaborative online AI engine for CT-based COVID-19 diagnosis," *MedRxiv*, 2020.

[98] A. Dofferhoff, A. Swinkels, T. Sprong, Y. Berk, M. Spanbroek, M. H. Nabuurs-Franssen, M. Vermaat, B. van de Kerkhof, M. H. Willekens, and A. Voss, "Diagnostic algorithm for COVID-19 at the ER," *Nederlands Tijdschrift voor Geneeskunde*, vol. 164, pp. D5042–D5042, 2020.

[99] S. Kurstjens, A. Van Der Horst, R. Herpers, M. W. Geerits, Y. C. Kluiters-de Hingh, E.-L. Göttgens, M. J. Blaauw, M. H. Thelen, M. G. Elisen, and R. Kusters, "Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 58, no. 9, pp. 1587–1593, 2020.

[100] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 424–432. [Online]. Available: https://doi.org/10.1007/978-3-319-46723-8_49

[101] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," *arXiv:2003.09424*, 2020.

[102] D. Singh, V. Kumar, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 39, no. 7, pp. 1–11, 2020.

[103] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, M. Zhang, L. Yu, F. Cheng, and H. Yu, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Scientific Reports*, vol. 10, no. 1, p. 19196, Nov 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76282-0

[104] Y. Li, D. Wei, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian *et al.*, "Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2787–2797, 2020.

[105] Y.-M. Xu, T. Zhang, H. Xu, L. Qi, W. Zhang, Y.-D. Zhang, D.-S. Gao, M. Yuan, and T.-F. Yu, "Deep learning in CT images: Automated pulmonary nodule detection for subsequent management using convolutional neural network," *Cancer Management and Research*, vol. 12, p. 2979, 2020.

[106] D. Di, F. Shi, F. Yan, L. Xia, Z. Mo, Z. Ding, F. Shan, B. Song, S. Li, Y. Wei *et al.*, "Hypergraph learning for identification of COVID-19 with CT imaging," *Medical Image Analysis*, vol. 68, p. 101910, 2020.

[107] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[108] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, 2013.

[109] B. van Ginneken, S. G. Armato, B. de Hoop, S. van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. M. R. Schilham, A. Retico, M. E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Belloti, F. D. Carlo, R. Megna, S. Tangaro, L. Bolanos, P. Cerello, S. C. Cheran, E. L. Torres, and M. Prokop, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Medical Image Analysis*, vol. 14, pp. 707–722, 12 2010.

[110] W. Ning, Y. Xue, S. Lei, L. Wang, Z. Wang, J. Yang, and Y. Zheng, "CT images and clinical features for COVID-19," http://ictcf.biocuckoo.cn/HUST-19.php, accessed: 2020-05-20.

[111] C. de Vente and L. H. Boulogne, "Grand Challenge - COVID-19 CT Classification challenge," https://covid19.grand-challenge.org/, accessed: 2020-09-01.

[112] I. W. Harsono, S. Liawatimena, and T. W. Cenggoro, "Lung nodule detection and classification from thorax CT-scan using RetinaNet with transfer learning," *Journal of King Saud University-Computer and Information Sciences*, 2020, to be published.

[113] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, "Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19," *Pattern Recognition*, vol. 114, p. 107848, 2021.

[114] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[115] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[116] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017, pp. 4278–4284.

[117] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[118] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[119] S. A. A. Ahmed, M. C. Yavuz, M. U. Sen, F. Gulsen, O. Tutar, B. Korkmazer, C. Samanci, S. Sirolu, R. Hamid, A. E. Eryurekli *et al.*, "COVID-19 detection in computed tomography images with 2D and 3D approaches," *arXiv preprint arXiv:2105.08506*, 2021.

[120] P. K. Chaudhary and R. B. Pachori, "FBSED based automatic diagnosis of COVID-19 using X-ray and CT images," *Computers in biology and medicine*, vol. 134, p. 104454, 2021.

[121] Z. Li, J. Zhang, B. Li, X. Gu, and X. Luo, "COVID-19 diagnosis on CT scan images using a generative adversarial network and concatenated feature pyramid network with an attention mechanism," *Medical Physics*, vol. 48, no. 8, pp. 4334–4349, 2021.

[122] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, 2019, pp. 3342–3352.

[123] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare *et al.*, "Automated detection and forecasting of COVID-19 using deep learning techniques: A review," *arXiv preprint arXiv:2007.10785*, 2020.

[124] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[125] C. de Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta, "Deep learning regression for prostate cancer detection and grading in bi-parametric MRI," *IEEE Transactions on Biomedical Engineering*, 2020.

[126] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.

[127] C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," *Academic Radiology*, vol. 7, pp. 413–419, 2000.

[128] F. Samuelson, N. Petrick, and S. Paquerault, "Advantages and examples of resampling for CAD evaluation," in *IEEE International Symposium on Biomedical Imaging*, 2007, pp. 492 –495.

[129] P. Mokhles, J. S. Schouten, H. J. Beckers, A. Azuara-Blanco, A. Tuulonen, and C. A. Webers, "A systematic review of end-of-life visual impairment in open-angle glaucoma: an epidemiological autopsy," *Journal of Glaucoma*, vol. 25, no. 7, pp. 623–628, 2016.

[130] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014.

[131] X. Chen, Y. Xu, F. Yin, Z. Zhang, D. W. K. Wong, T. Y. Wong, and J. Liu, "Multiple ocular diseases detection based on joint sparse multi-task learning," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 5260–5263.

[132] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes." *Journal of the American Medical Association*, vol. 318, pp. 2211–2223, Dec. 2017.

[133] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.

[134] S. Phene, R. C. Dunn, N. Hammel, Y. Liu, J. Krause, N. Kitade, M. Schaekermann, R. Sayres, D. J. Wu, A. Bora *et al.*, "Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs," *Ophthalmology*, vol. 126, no. 12, pp. 1627–1639, 2019.

[135] T. W. Rogers, N. Jaccard, F. Carbonaro, H. G. Lemij, K. A. Vermeer, N. J. Reus, and S. Trikha, "Evaluation of an ai system for the automated detection of glaucoma from stereoscopic optic disc photographs: the european optic disc assessment study," *Eye*, vol. 33, no. 11, pp. 1791–1797, 2019.

[136] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 3065–3068.

[137] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2011, pp. 1–6.

[138] J. Odstrcilik, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, and E. Angelopoulou, "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database," *IET Image Processing*, vol. 7, no. 4, pp. 373–383, 2013.

[139] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[140] S. Holm, G. Russell, V. Nourrit, and N. McLoughlin, "Dr hagis—a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients," *Journal of Medical Imaging*, vol. 4, no. 1, p. 014503, 2017.

[141] J. I. Orlando, J. Barbosa Breda, K. v. Keer, M. B. Blaschko, P. J. Blanco, and C. A. Bulant, "Towards a glaucoma risk index based on simulated hemodynamics from fundus images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 65–73.

[142] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[143] H. Fang, F. Li, H. Fu, X. Sun, X. Cao, J. Son, S. Yu, M. Zhang, C. Yuan, C. Bian *et al.*, "Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment," *arXiv preprint arXiv:2202.08994*, 2022.

[144] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, L. Huang, Q. Yu, S. Song, X. Xu *et al.*, "Gamma challenge: glaucoma grading from multi-modality images," *arXiv preprint arXiv:2202.06511*, 2022.

[145] J. Cuadros and G. Bresnick, "Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening," *Journal of diabetes science and technology*, vol. 3, no. 3, pp. 509–516, 2009.

[146] N. J. Reus, H. G. Lemij, D. F. Garway-Heath, P. J. Airaksinen, A. Anton, A. M. Bron, C. Faschinger, G. Holló, M. Iester, J. B. Jonas *et al.*, "Clinical assessment of stereoscopic optic disc photographs for glaucoma: the european optic disc assessment trial," *Ophthalmology*, vol. 117, no. 4, pp. 717–723, 2010.

[147] H. G. Lemij, C. de Vente, C. I. Sánchez, and K. A. Vermeer, "Characteristics of a large, labeled dataset for the training of artificial intelligence for glaucoma screening with fundus photographs," *Ophthalmology Science*, 2023.

[148] U. Sevik, C. Kose, T. Berber, and H. Erdol, "Identification of suitable fundus images using automated quality assessment methods," *Journal of biomedical optics*, vol. 19, no. 4, p. 046006, 2014.

[149] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical decision making*, vol. 9, no. 3, pp. 190–195, 1989.

[150] H. Vaahtoranta-Lehtonen, A. Tuulonen, P. Aronen, H. Sintonen, L. Suoranta, N. Kovanen, M. Linna, E. Läärä, and A. Malmivaara, "Cost effectiveness and cost utility of an organized screening programme for glaucoma," *Acta Ophthalmologica Scandinavica*, vol. 85, no. 5, pp. 508–518, 2007.

[151] M. M. de Vries, R. Stoutenbeek, R. P. Müskens, and N. M. Jansonius, "Glaucoma screening during regular optician visits: the feasibility and specificity of screening in real life," *Acta ophthalmologica*, vol. 90, no. 2, pp. 115–121, 2012.

[152] C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," *Academic radiology*, vol. 7, no. 6, pp. 413–419, 2000.

[153] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[154] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[155] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[156] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[157] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[158] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[159] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[160] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 774–782.

[161] H. Wang, H. Sun, Y. Fang, S. Li, M. Feng, and R. Wang, "A workflow for computer-aided diagnosis of glaucoma," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–4.

[162] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[163] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" *arXiv preprint arXiv:2103.13915*, 2021.

[164] W. Liu, H. Yang, T. Tian, Z. Cao, X. Pan, W. Xu, Y. Jin, and F. Gao, "Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4623–4634, 2022.

[165] F. Khader, C. Haarburger, J.-C. Kirr, M. Menke, J. N. Kather, J. Stegmaier, C. Kuhl, S. Nebelung, and D. Truhn, "Elevating fundoscopic evaluation to expert level-automatic glaucoma detection using data from the airogs challenge," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–4.

[166] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[167] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[168] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[169] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, F. Ingham, J. Poznanski, J. Fang, L. Yu *et al.*, "ultralytics/yolov5," *Zenodo*, 2020.

[170] T. Aimyshev and K. Eleusiz. (2022) Glaucoma detection algorithm for the artificial intelligence for robust glaucoma screening challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/34b337f9-551a-4c82-834a-32a9f99ba690/AIROGS.pdf

[171] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[172] T.-D. Le. (2022) Combination of supervised learning and unsupervised learning to detect ungradable images in the airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/3e9f9bbb-9d22-430f-9448-6a4e21a29f2f/Airogs_Combination_Algorithm.pdf

[173] A. Galdran, G. Carneiro, and M. A. G. Ballester. (2022) Open-set glaucoma screening from eye fundus images: Domain knowledge to the rescue. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/e01b1325-be85-4d1e-9322-c0c3726ab9ae/airogs_osr.pdf

[174] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[175] D. Puthussery, P. S. Hrishikesh, R. G. Devika, and C. V. Jiji. (2022) A self-supervised approach for glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/fd5a5254-8faf-42df-97ec-b500c35c7c2a/A_Self-Supervised_appr_GsyqCdA.pdf

[176] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[177] P. Oza and V. M. Patel, "One-class convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2018.

[178] Z. Yang, H. Liu, and Z. Shang. (2022) Deep learning for referable glaucoma screening and out-of-distribution detection. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/d259c884-5eda-4bb9-b9fb-73a29e9eaf41/airogs.pdf

[179] S. Kondo, S. Kasai, and K. Hirasawa, "Computer aided diagnosis and out-of-distribution detection in glaucoma screening using color fundus photography," *arXiv preprint arXiv:2202.11944*, 2022.

[180] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph, "Revisiting ResNets: Improved training and scaling strategies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 614–22 627, 2021.

[181] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.

[182] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[183] E. Wang, A. Durvasula, D. Deng, A. Sivajohan, E. Ho, and K. Lane. (2022) Ensemble network for glaucoma screening in airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/ea161fad-ca7f-4e1a-9133-bed2128107c3/AIROGS_Manuscript.pdf

[184] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[185] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[186] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.

[187] J. Heras, D. Royo, and M. A. Zapata. (2022) A good closed-set classifier is all you need for the airogs challenge. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/d79d3e55-505a-416a-b389-0a51170b1271/AIROGS.pdf

[188] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," *arXiv preprint arXiv:2110.06207*, 2021.

[189] T. Araújo, G. Aresta, and H. Bogunović, "Deep dirichlet uncertainty for unsupervised out-of-distribution detection of eye fundus photographs in glaucoma screening," in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022, pp. 1–5.

[190] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[191] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[192] M. Arikan. (2022) Multi-model ensemble for robust glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/c9b4a350-62f4-4f80-aa16-cfc71b6640c3/Multi_model_ensemble_f_dsy4aS0.pdf

[193] Y. C. Lee, H. B. Cho, and Y. H. Choi. (2022) Classification for referable glaucoma with fundus photographs using multimodal deep learning. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/7b0a4c21-e8b4-4fee-9e01-f48e25b2b1b4/Classification_for_ref_sFDsTWh.pdf

[194] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016, pp. 1050–1059.

[195] A. Qayyum, M. Mazher, and I. Razzak. (2022) ConvNeXts and vision transformer based framework for glaucoma screening. [Online]. Available: https://rumc-gcorg-p-public.s3.amazonaws.com/evaluation-supplementary/644/33dea053-4ae5-4ac4-8772-f4599c6af590/ISIB_AIROGS__Imran.pdf

[196] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[197] K. Bibas, M. Feder, and T. Hassner, "Single layer predictive normalized maximum likelihood for out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1179–1191, 2021.

[198] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[199] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical image analysis*, p. 102762, 2023.

[200] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3D medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 730–20 740.

[201] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

[202] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[203] C. González-Gonzalo, E. F. Thee, C. C. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy AI: Closing the gap between development and integration of ai systems in ophthalmic practice," *Progress in retinal and eye research*, vol. 90, p. 101034, 2022.

[204] R. T. Yanagihara, C. S. Lee, D. S. W. Ting, and A. Y. Lee, "Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 11–11, 2020.

[205] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[206] C. González-Gonzalo, E. F. Thee, C. C. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice," *Progress in Retinal and Eye Research*, p. 101034, 2021.

[207] European Commission. (2019) Ethics guidelines for trustworthy AI. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[208] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *arXiv preprint arXiv:1906.02530*, 2019.

[209] S. Neppalli, M. A. Kessell, C. R. Madeley, M. L. Hill, P. S. Vlaskovsky, and D. B. Taylor, "Artifacts in contrast-enhanced mammography: are there differences between vendors?" *Clinical Imaging*, vol. 80, pp. 123–130, 2021.

[210] F. Bazvand and F. Ghassemi, "Artifacts in macular optical coherence tomography," *Journal of Current Ophthalmology*, vol. 32, no. 2, p. 123, 2020.

[211] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[212] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2127–2136.

[213] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[214] O. Puzyeyeva, W. C. Lam, J. G. Flanagan, M. H. Brent, R. G. Devenyi, M. S. Mandelcorn, T. Wong, C. Hudson *et al.*, "High-resolution optical coherence tomography retinal imaging: a case series illustrating potential and limitations," *Journal of ophthalmology*, vol. 2011, 2011.

[215] J. Linmans, S. Elfwing, J. van der Laak, and G. Litjens, "Predictive uncertainty estimation for out-of-distribution detection in digital pathology," *Medical Image Analysis*, vol. 83, p. 102655, 2023.

[216] B. Schömig-Markiefka, A. Pryalukhin, W. Hulla, A. Bychkov, J. Fukuoka, A. Madabhushi, V. Achter, L. Nieroda, R. Büttner, A. Quaas *et al.*, "Quality control stress test for deep learning-based diagnostic model in digital pathology," *Modern Pathology*, vol. 34, no. 12, pp. 2098–2108, 2021.

[217] R. Wellenberg, E. Hakvoort, C. Slump, M. Boomsma, M. Maas, and G. Streekstra, "Metal artifact reduction techniques in musculoskeletal CT-imaging," *European Journal of Radiology*, vol. 107, pp. 60–69, 2018.

[218] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "Camel: A weakly supervised learning framework for histopathology image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 682–10 691.

[219] A. Patil, D. Tamboli, S. Meena, D. Anand, and A. Sethi, "Breast cancer histopathology image classification and localization using multiple instance learning," in *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2019, pp. 1–4.

[220] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 519–528.

[221] J. M. Tomczak, M. Ilse, M. Welling, M. Jansen, H. G. Coleman, M. Lucas, K. de Laat, M. de Bruin, H. Marquering, M. J. van der Wel *et al.*, "Histopathological classification of precursor lesions of esophageal adenocarcinoma: A deep multiple instance learning approach," in *Medical Imaging with Deep Learning*, 2018, pp. 1–3.

[222] S. Yin, Q. Peng, H. Li, Z. Zhang, X. You, H. Liu, K. Fischer, S. L. Furth, G. E. Tasian, and Y. Fan, "Multi-instance deep learning with graph convolutional neural networks for diagnosis of kidney diseases using ultrasound imaging," in *Uncertainty for safe utilization of machine learning in medical imaging and clinical image-based procedures*. Springer, 2019, pp. 146–154.

[223] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 38, no. 3, pp. 762–774, 2018.

[224] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang, "Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2584–2594, 2020.

[225] C. Xu, S. Qi, J. Feng, S. Xia, Y. Kang, Y. Yao, and W. Qian, "DCT-MIL: deep CNN transferred multiple instance learning for COPD identification using CT images," *Physics in Medicine & Biology*, vol. 65, no. 14, p. 145011, 2020.

[226] W. Zhu, L. Sun, J. Huang, L. Han, and D. Zhang, "Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2354–2366, 2021.

[227] Z. Qiu, Y. Pan, J. Wei, D. Wu, Y. Xia, and D. Shen, "Predicting symptoms from multiphasic MRI via multi-instance attention learning for hepatocellular carcinoma grading," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 439–448.

[228] Q. Ren, Y. Zhao, B. He, B. Wu, S. Mai, F. Xu, Y. Huang, Y. He, J. Huang, and J. Yao, "IIB-MIL: Integrated instance-level and bag-level multiple instances learning with label disambiguation for pathological image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 560–569.

[229] L. Qu, M. Wang, Z. Song *et al.*, "Bi-directional weakly supervised knowledge distillation for whole slide image classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 368–15 381, 2022.

[230] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[231] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 951–10 960.

[232] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty detection via contrastive learning on distributionally shifted instances," *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.

[233] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[234] E. Calli, E. Sogancioglu, E. T. Scholten, K. Murphy, and B. van Ginneken, "Handling label noise through model confidence and uncertainty: application to chest radiograph classification," in *Medical Imaging*, ser. Proceedings of the SPIE, no. 1, 3 2019.

[235] J. Linmans, J. van der Laak, and G. Litjens, "Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 465–478. [Online]. Available: https://openreview.net/forum?id=hRwB2BTRNu

[236] S. Apostolopoulos, C. Ciller, S. De Zanet, S. Wolf, and R. Sznitman, "RetiNet: Automatic AMD identification in OCT volumetric data," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 8, pp. 387–387, 2017.

[237] F. G. Venhuizen, B. van Ginneken, F. van Asten, M. J. van Grinsven, S. Fauser, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Automated staging of age-related macular degeneration using optical coherence tomography," *Investigative Ophthalmology and Visual Science*, vol. 58, no. 4, pp. 2318–2328, 4 2017.

[238] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," *IEEE Transactions on Medical Imaging*, vol. 37, no. 4, pp. 1024–1034, 2017.

[239] T. Kurmann, S. Yu, P. Márquez-Neila, A. Ebneter, M. Zinkernagel, M. R. Munk, S. Wolf, and R. Sznitman, "Expert-level automated biomarker identification in optical coherence tomography scans," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[240] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration OCT images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[241] X. Wang, F. Tang, H. Chen, L. Luo, Z. Tang, A.-R. Ran, C. Y. Cheung, and P.-A. Heng, "UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3431–3442, 2020.

[242] P. Seeböck, D. Romo-Bucheli, S. Waldstein, H. Bogunovic, J. I. Orlando, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Using CycleGANs for effectively reducing image variability across OCT devices and improving retinal fluid segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 605–609.

[243] D. Romo-Bucheli, P. Seeböck, J. I. Orlando, B. S. Gerendas, S. M. Waldstein, U. Schmidt-Erfurth, and H. Bogunović, "Reducing image variability across OCT devices with unsupervised unpaired learning for improved segmentation of retina," *Biomedical Optics Express*, vol. 11, no. 1, pp. 346–363, 2020.

[244] M. U. Oner, J. M. S. Kye-Jet, H. K. Lee, and W.-K. Sung, "Distribution based mil pooling filters: Experiments on a lymph node metastases dataset," *Medical Image Analysis*, vol. 87, p. 102813, 2023.

[245] ——, "Studying the effect of mil pooling filters on mil tasks," *arXiv preprint arXiv:2006.01561*, 2020.

[246] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[247] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens, "Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection," *Medical image analysis*, vol. 64, p. 101724, 2020.

[248] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*, 2019.

[249] J. P. H. van de Ven, D. Smailhodzic, C. J. F. Boon, S. Fauser, J. M. M. Groenewoud, N. Victor Chong, C. B. Hoyng, B. Jeroen Klevering, and A. I. den Hollander, "Association analysis of genetic and environmental risk factors in the cuticular drusen subtype of age-related macular degeneration," *Molecular Vision*, vol. 18, pp. 2271–2278, 2012.

[250] S. Fauser, D. Smailhodzic, A. Caramoy, J. P. H. van de Ven, B. Kirchhof, C. B. Hoyng, B. Jeroen Klevering, S. Liakopoulos, and A. I. den Hollander, "Evaluation of serum lipid concentrations and genetic variants at high-density lipoprotein metabolism loci and TIMP3 in age-related macular degeneration," *Investigative Ophthalmology and Visual Science*, vol. 52, no. 8, pp. 5525–5528, 2011.

[251] M. A. Ikram, G. G. Brusselle, S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure, C. C. Klaver, T. E. Nijsten, R. P. Peeters, B. H. Stricker *et al.*, "The Rotterdam Study: 2018 update on objectives, design and main results," *European Journal of Epidemiology*, vol. 32, no. 9, pp. 807–850, 2017.

[252] R. Klein, M. D. Davis, Y. L. Magli, P. Segal, B. E. K. Klein, and L. Hubbard, "The Wisconsin age-related maculopathy grading system," *Ophthalmology*, vol. 98, no. 7, pp. 1128–1134, 1991.

[253] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[254] E. Y. Chew, T. Clemons, J. P. SanGiovanni, R. Danis, A. Domalpally, W. McBee, R. Sperduto, F. L. Ferris, A. R. Group *et al.*, "The age-related eye disease study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1)," *Ophthalmology*, vol. 119, no. 11, pp. 2282–2289, 2012.

[255] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.

[256] J. Ho, D. P. Castro, L. C. Castro, Y. Chen, J. Liu, C. Mattox, C. Krishnan, J. G. Fujimoto, J. S. Schuman, and J. S. Duker, "Clinical assessment of mirror artifacts in spectral-domain optical coherence tomography," *Investigative Ophthalmology & Visual Science*, vol. 51, no. 7, pp. 3714–3720, 2010.

[257] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[258] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[259] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[260] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[261] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[262] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

[263] N. Kanwal, M. López-Pérez, U. Kiraz, T. C. Zuiverloon, R. Molina, and K. Engan, "Are you sure it's an artifact? artifact detection and uncertainty quantification in histological images," *Computerized Medical Imaging and Graphics*, vol. 112, p. 102321, 2024.

[264] A. Foucart, O. Debeir, and C. Decaestecker, "Artifact identification in digital pathology from weak and noisy supervision with deep residual networks," in *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech).* IEEE, 2018, pp. 1–6.

[265] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[266] R. P. Finger, S. Schmitz-Valckenberg, M. Schmid, G. S. Rubin, H. Dunbar, A. Tufail, D. P. Crabb, A. Binns, C. I. Sanchez, P. Margaron, G. Normand, M. K. Durbin, U. F. O. Luhmann, P. Zamiri, J. Cunha-Vaz, F. Asmus, F. G. Holz, and on behalf of the MACUSTAR consortium, "Macustar: Development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration," *Ophthalmologica*, vol. 241, pp. 61–72, 8 2019. [Online]. Available: https://www.karger.com/Article/FullText/491402

[267] L. de Sisternes, N. Simon, R. Tibshirani, T. Leng, and D. L. Rubin, "Quantitative sd-oct imaging biomarkers as indicators of age-related macular degeneration progression," *Investigative Opthalmology & Visual Science*, vol. 55, no. 11, p. 7093, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1167/iovs.14-14918

[268] Z. Wu, C. D. Luu, L. N. Ayton, J. K. Goh, L. M. Lucci, W. C. Hubbard, J. L. Hageman, G. S. Hageman, and R. H. Guymer, "Optical coherence tomography–defined changes preceding the development of drusen-associated atrophy in age-related macular degeneration," *Ophthalmology*, vol. 121, no. 12, p. 2415–2422, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ophtha.2014.06.034

[269] M. Fleckenstein, T. D. L. Keenan, R. H. Guymer, U. Chakravarthy, S. Schmitz-Valckenberg, C. C. Klaver, W. T. Wong, and E. Y. Chew, "Age-related macular degeneration," *Nature Reviews Disease Primers*, vol. 7, no. 1, May 2021. [Online]. Available: http://dx.doi.org/10.1038/s41572-021-00265-2

[270] E. M. Lad, R. P. Finger, and R. Guymer, "Biomarkers for the progression of intermediate age-related macular degeneration," *Ophthalmology and Therapy*, vol. 12, no. 6, p. 2917–2941, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1007/s40123-023-00807-9

[271] Z. Wu, E. L. Fletcher, H. Kumar, U. Greferath, and R. H. Guymer, "Reticular pseudodrusen: A critical phenotype in age-related macular degeneration," *Progress in Retinal and Eye Research*, vol. 88, p. 101017, May 2022. [Online]. Available: http://dx.doi.org/10.1016/j.preteyeres.2021.101017

[272] J. G. Christenbury, F. A. Folgar, R. V. O'Connell, S. J. Chiu, S. Farsiu, and C. A. Toth, "Progression of intermediate age-related macular degeneration with proliferation and inner retinal migration of hyperreflective foci," *Ophthalmology*, vol. 120, no. 5, p. 1038–1045, May 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ophtha.2012.10.018

[273] M. Veerappan, A.-K. M. El-Hage-Sleiman, V. Tai, S. J. Chiu, K. P. Winter, S. S. Stinnett, T. S. Hwang, G. B. Hubbard, M. Michelson, R. Gunther, W. T. Wong, E. Y. Chew, C. A. Toth, C. A. Toth, W. Wong, T. Hwang, G. B. Hubbard, S. Srivastava, M. McCall, K. Winter, N. Sarin, K. Hall, P. McCollum, L. Curtis, S. Schuman, S. J. Chiu, S. Farsiu, V. Tai, M. Sevilla, C. Harrington, R. Gunther, D. Tran-Viet, F. Folgar, E. Yuan, T. Clemons, M. Harrington, and E. Chew, "Optical coherence tomography reflective drusen substructures predict progression to geographic atrophy in age-related macular degeneration," *Ophthalmology*, vol. 123, no. 12, p. 2554–2570, Dec. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ophtha.2016.08.047

[274] S. Thiele, J. Nadal, M. Pfau, M. Saßmannshausen, M. Fleckenstein, F. G. Holz, M. Schmid, and S. Schmitz-Valckenberg, "Prognostic value of intermediate age-related macular degeneration phenotypes for geographic atrophy progression," *British Journal of Ophthalmology*, vol. 105, no. 2, p. 239–245, Apr. 2020. [Online]. Available: http://dx.doi.org/10.1136/bjophthalmol-2020-316004

[275] A. M. Hagag, R. Kaye, V. Hoang, S. Riedl, P. Anders, B. Stuart, G. Traber, C. Appenzeller-Herzog, U. Schmidt-Erfurth, H. Bogunovic, H. P. Scholl, T. Prevost, L. Fritsche, D. Rueckert, S. Sivaprasad, and A. J. Lotery, "Systematic review of prognostic factors associated with progression to late age-related macular degeneration: Pinnacle study report 2," *Survey of Ophthalmology*, vol. 69, no. 2, p. 165–172, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.survophthal.2023.10.010

[276] M. Nassisi, J. Lei, N. S. Abdelfattah, A. Karamat, S. Balasubramanian, W. Fan, A. Uji, K. M. Marion, K. Baker, X. Huang, E. Morgenthien, and S. R. Sadda, "OCT risk factors for development of late age-related macular degeneration in the fellow eyes of patients enrolled in the HARBOR study," *Ophthalmology*, vol. 126, no. 12, p. 1667–1674, Dec. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.ophtha.2019.05.016

[277] J. Sutton, M. J. Menten, S. Riedl, H. Bogunović, O. Leingang, P. Anders, A. M. Hagag, S. Waldstein, A. Wilson, A. J. Cree, G. Traber, L. G. Fritsche, H. Scholl, D. Rueckert, U. Schmidt-Erfurth, S. Sivaprasad, T. Prevost, and A. Lotery, "Developing and validating a multivariable prediction model which predicts progression of intermediate to late age-related macular degeneration—the pinnacle trial protocol," *Eye*, vol. 37, no. 6, p. 1275–1283, May 2022. [Online]. Available: http://dx.doi.org/10.1038/s41433-022-02097-0

[278] J. Lei, S. Balasubramanian, N. S. Abdelfattah, M. G. Nittala, and S. R. Sadda, "Proposal of a simple optical coherence tomography-based scoring system for progression of age-related macular degeneration," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 255, no. 8, p. 1551–1558, May 2017. [Online]. Available: http://dx.doi.org/10.1007/s00417-017-3693-y

[279] S. R. Sadda, R. Guymer, F. G. Holz, S. Schmitz-Valckenberg, C. A. Curcio, A. C. Bird, B. A. Blodi, F. Bottoni, U. Chakravarthy, E. Y. Chew, K. Csaky, R. P. Danis, M. Fleckenstein, K. B. Freund, J. Grunwald, C. B. Hoyng, G. J. Jaffe, S. Liakopoulos, J. M. Monés, D. Pauleikhoff, P. J. Rosenfeld, D. Sarraf, R. F. Spaide, R. Tadayoni, A. Tufail, S. Wolf, and G. Staurenghi, "Consensus definition for atrophy associated with age-related macular degeneration on oct," *Ophthalmology*, vol. 125, no. 4, p. 537–548, Apr. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.ophtha.2017.09.028

[280] R. H. Guymer, Z. Wu, S. Gao, M. Zhang, V. Steffen, K. Litts, M. C. Chang, J. Elsrott, C. Chen, J. Zhang *et al.*, "Honu: A multicenter, prospective, observational study of the progression of intermediate age-related macular degeneration," *Investigative Ophthalmology & Visual Science*, vol. 64, no. 8, pp. 2754–2754, 2023.

[281] G. Corradetti, F. Corvi, M. G. Nittala, M. Nassisi, A. R. Alagorie, J. Scharf, M. Y. Lee, S. R. Sadda, and D. Sarraf, "Natural history of incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration," *Canadian Journal of Ophthalmology*, vol. 56, no. 5, p. 325–334, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.jcjo.2021.01.005

[282] J. N. Chiang, G. Corradetti, M. G. Nittala, F. Corvi, N. Rakocz, A. Rudas, B. Durmus, U. An, S. Sankararaman, A. Chiu, E. Halperin, and S. R. Sadda, "Automated identification of incomplete and complete retinal epithelial pigment and outer retinal atrophy using machine learning," *Ophthalmology Retina*, vol. 7, no. 2, p. 118–126, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.oret.2022.08.016

[283] A. Szeskin, R. Yehuda, O. Shmueli, J. Levy, and L. Joskowicz, "A column-based deep learning method for the detection and quantification of atrophy associated with AMD in OCT scans," *Medical Image Analysis*, vol. 72, p. 102130, Aug. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.media.2021.102130

[284] G. Zhang, D. J. Fu, B. Liefers, L. Faes, S. Glinton, S. Wagner, R. Struyven, N. Pontikos, P. A. Keane, and K. Balaskas, "Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study," *The Lancet Digital Health*, vol. 3, no. 10, p. e665–e675, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1016/S2589-7500(21)00134-5

[285] C. de Vente, B. van Ginneken, C. B. Hoyng, C. C. Klaver, and C. I. Sánchez, "Uncertainty-aware multiple-instance learning for reliable classification: Application to optical coherence tomography," *Medical Image Analysis*, vol. 97, p. 103259, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524001841

[286] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, J. Kulisevsky, F. Blanc, A. Antonini, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, D. Aarsland, and E. Westman, "The reliability of a deep learning model in clinical out-of-distribution mri data: A multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.media.2020.101714

[287] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, "Data drift in medical machine learning: implications and potential remedies," *The British Journal of Radiology*, vol. 96, no. 1150, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1259/bjr.20220878

[288] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, Nov. 2018. [Online]. Available: http://dx.doi.org/10.1371/journal.pmed.1002683

[289] E. H. P. Pooch, P. Ballester, and R. C. Barros, *Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification*. Springer International Publishing, 2020, p. 74–83. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-62469-9_7

[290] J. H. Terheyden, F. G. Holz, S. Schmitz-Valckenberg, A. Lüning, M. Schmid, G. S. Rubin, H. Dunbar, A. Tufail, D. P. Crabb, A. Binns, C. I. Sánchez, C. Hoyng, P. Margaron, N. Zakaria, M. Durbin, U. Luhmann, P. Zamiri, J. Cunha-Vaz, C. Martinho, S. Leal, R. P. Finger, P. Basile, C. Behning, M. Berger, A. Binns, M. Böttger, C. Bouchet, J. E. Brazier, T. Butt, C. Carapezzi, J. Carlton, A. Charil, R. Coimbra, S. Nunes, D. Crabb, J. Cunha-Vaz, H. Dunbar, M. Durbin, R. Finger, F. Holz, C. Hoyng, J. Krätzschmar, S. Leal, U. Luhmann, A. Lüning, P. Margaron, C. Martinho, B. Melício, S. Mohand-Saïd, D. Rowen, G. S. Rubin, J. Sahel, C. I. Sánchez, D. Sanches Fernandes, M. Schmid, S. Schmitz-Valckenberg, A. Skelly, L. Stöhr, D. Taylor, J. Terheyden, A. Tufail, L. Vieweg, L. Wintergerst, C. Wojek, N. Zakaria, and P. Zamiri, "Clinical study protocol for a low-interventional study in intermediate age-related macular degeneration developing novel clinical endpoints for interventional clinical trials with a regulatory and patient access intention—macustar," *Trials*, vol. 21, no. 1, Jul. 2020. [Online]. Available: http://dx.doi.org/10.1186/s13063-020-04595-6

[291] M. Saßmannshausen, C. Behning, J. Weinz, L. Goerdt, J. H. Terheyden, P. Chang, M. Schmid, S. H. Poor, N. Zakaria, R. P. Finger, F. G. Holz, M. Pfau, S. Schmitz-Valckenberg, S. Thiele, H. Agostini, L. Altay, R. Atia, F. Bandello, P. Basile, C. Behning, M. Belmouhand, M. Berger, A. Binns, C. Boon, M. Böttger, C. Bouchet, J. Brazier, T. Butt, C. Carapezzi, J. Carlton, A. Carneiro, A. Charil, R. Coimbra, M. Cozzi, D. Crabb, J. Cunha-Vaz, C. Dahlke, L. de Sisternes, H. Dunbar, R. Finger, E. Fletcher, H. Floyd, C. Francisco, M. Gutfleisch, R. Hogg, F. Holz, C. Hoyng, A. Kilani, J. Krätzschmar, L. Kühlewein, M. Larsen, S. Leal, Y. Lechanteur, U. Luhmann, A. Lüning, I. Marques, C. Martinho, G. Montesano, Z. Mulyukov, M. Paques, B. Parodi, M. Parravano, S. Penas, T. Peters, T. Peto, M. Pfau, S. Poor, S. Priglinger, D. Rowen, G. Rubin, J. Sahel, C. Sánchez, O. Sander, M. Saßmannshausen, M. Schmid, S. Schmitz-Valckenberg, H. Schrinner-Fenske, J. Siedlecki, R. Silva, A. Skelly, E. Souied, G. Staurenghi, L. Stöhr, D. Taylor, J. Terheyden, S. Thiele, A. Tufail, M. Varano, L. Vieweg, L. Wintergerst, A. Wolf, and N. Zakaria, "Characteristics and spatial distribution of structural features in age-related macular degeneration," *Ophthalmology Retina*, vol. 7, no. 5, p. 420–430, May 2023. [Online]. Available: http://dx.doi.org/10.1016/j.oret.2022.12.007

[292] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[293] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[294] M. J. Mulder, M. C. Keuken, P.-L. Bazin, A. Alkemade, and B. U. Forstmann, "Size and shape matter: The impact of voxel geometry on the identification of small nuclei," *PLOS ONE*, vol. 14, no. 4, pp. 1–19, 04 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0215382

[295] J. Wu, H. Fang, F. Li, H. Fu, F. Lin, J. Li, Y. Huang, Q. Yu, S. Song, X. Xu *et al.*, "Gamma challenge: glaucoma grading from multi-modality images," *Medical Image Analysis*, vol. 90, p. 102938, 2023.

[296] D. Garzone, J. H. Terheyden, O. Morelle, M. W. Wintergerst, M. Saßmannshausen, S. Schmitz-Valckenberg, M. Pfau, S. Thiele, S. Poor, S. Leal *et al.*, "Comparability of automated drusen volume measurements in age-related macular degeneration: a macustar study report," *Scientific reports*, vol. 12, no. 1, p. 21911, 2022.

[297] S. Velaga, M. Nittala, R. Konduru, F. Heussen, P. Keane, and S. Sadda, "Impact of optical coherence tomography scanning density on quantitative analyses in neovascular age-related macular degeneration," *Eye*, vol. 31, no. 1, pp. 53–61, 2017.

[298] A. E. Barañano, P. A. Keane, H. Ruiz-Garcia, A. C. Walsh, and S. R. Sadda, "Impact of scanning density on spectral domain optical coherence tomography assessments in neovascular age-related macular degeneration," *Acta Ophthalmologica*, vol. 90, no. 4, pp. e274–e280, 2012.

[299] D. Oncel, N. Manafi, M. G. Nittala, S. B. Velaga, D. Stambolian, M. A. Pericak-Vance, J. L. Haines, and S. R. Sadda, "Effect of OCT B-scan density on sensitivity for detection of intraretinal hyperreflective foci in eyes with age-related macular degeneration," *Current eye research*, vol. 47, no. 9, pp. 1294–1299, 2022.

[300] M. G. Nittala, R. Konduru, H. Ruiz-Garcia, and S. Sadda, "Effect of OCT volume scan density on thickness measurements in diabetic macular edema," *Eye*, vol. 25, no. 10, pp. 1347–1355, 2011.

[301] Y. Huang, Z. Lu, Z. Shao, M. Ran, J. Zhou, L. Fang, and Y. Zhang, "Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network," *Optics express*, vol. 27, no. 9, pp. 12 289–12 307, 2019.

[302] Y. Xu, B. M. Williams, B. Al-Bander, Z. Yan, Y.-c. Shen, and Y. Zheng, "Improving the resolution of retinal oct with deep learning," in *Medical Image Understanding and Analysis: 22nd Conference, MIUA 2018, Southampton, UK, July 9-11, 2018, Proceedings 22*. Springer, 2018, pp. 325–332.

[303] V. Das, S. Dandapat, and P. K. Bora, "Unsupervised super-resolution of oct images using generative adversarial network for improved age-related macular degeneration diagnosis," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8746–8756, 2020.

[304] B. Qiu, Y. You, Z. Huang, X. Meng, Z. Jiang, C. Zhou, G. Liu, K. Yang, Q. Ren, and Y. Lu, "N2NSR-OCT: simultaneous denoising and super-resolution in optical coherence tomography images using semisupervised deep learning," *Journal of biophotonics*, vol. 14, no. 1, p. e202000282, 2021.

[305] J. Li, J. C. Koh, and W.-S. Lee, "HRINet: alternative supervision network for high-resolution CT image interpolation," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1916–1920.

[306] J. Sander, B. D. de Vos, and I. Išgum, "Autoencoding low-resolution MRI for semantically smooth interpolation of anisotropic MRI," *Medical image analysis*, vol. 78, p. 102393, 2022.

[307] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[308] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[309] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[310] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.

[311] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[312] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[313] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[314] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[315] W. H. Pinaya, M. S. Graham, E. Kerfoot, P.-D. Tudosiu, J. Dafflon, V. Fernandez, P. Sanchez, J. Wolleb, P. F. da Costa, A. Patel *et al.*, "Generative AI for medical imaging: extending the MONAI framework," *arXiv preprint arXiv:2307.15208*, 2023.

[316] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[317] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[318] S. Kim, C. Jin, T. Diethe, M. Figini, H. F. Tregidgo, A. Mullokandov, P. Teare, and D. C. Alexander, "Tackling structural hallucination in image translation with local diffusion," *arXiv preprint arXiv:2404.05980*, 2024.

[319] J. H. Terheyden, F. G. Holz, S. Schmitz-Valckenberg, A. Lüning, M. Schmid, G. S. Rubin, H. Dunbar, A. Tufail, D. P. Crabb, A. Binns *et al.*, "Clinical study protocol for a low-interventional study in intermediate age-related macular degeneration developing novel clinical endpoints for interventional clinical trials with a regulatory and patient access intention—macustar," *Trials*, vol. 21, pp. 1–11, 2020.

[320] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[321] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[322] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[323] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[324] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[325] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, pp. 69–101, 1996.

[326] J. M. Burr, G. Mowatt, R. Hernández, M. A. R. Siddiqui, J. Cook, T. Lourenco, C. Ramsay, L. Vale, C. Fraser, A. Azuara-Blanco *et al.*, "The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation." *Health technology assessment (Winchester, England)*, vol. 11, no. 41, pp. iii–190, 2007.

[327] R. Chou, S. Selph, I. Blazina, C. Bougatsos, R. Jungbauer, R. Fu, S. Grusing, D. E. Jonas, and S. Tehrani, "Screening for glaucoma in adults: updated evidence report and systematic review for the US preventive services task force," *JAMA*, vol. 327, no. 20, pp. 1998–2012, 2022.

[328] S. J. Gedde, K. Vinod, M. M. Wright, K. W. Muir, J. T. Lind, P. P. Chen, T. Li, and S. L. Mansberger, "Primary open-angle glaucoma preferred practice pattern®," *Ophthalmology*, vol. 128, no. 1, pp. P71–P150, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642020310241

[329] R. Tan, K. Y. C. Teo, R. Husain, N. C. Tan, Q. X. Lee, H. Hamzah, T. Wong, T. Aung, C. Y. Cheng, E. L. Lamoureux *et al.*, "Evaluating the outcome of screening for glaucoma using colour fundus photography-based referral criteria in a teleophthalmology screening programme for diabetic retinopathy," *British Journal of Ophthalmology*, vol. 108, no. 7, pp. 933–939, 2024.

[330] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

Summary

Deep learning (DL) models for medical image analysis have gained substantial popularity over the past decade. Despite high performance in lab settings, these models often fall short when applied in the real world. This can have considerable negative impact on both trust in these algorithms and their clinical use. A major cause of this failure is the absence of intrinsic mechanisms within most systems to enhance robustness. In this thesis, we addressed three facets of robustness: (1) the model's ability to perform well on data very close to or identical to the training distribution, (2) the model's ability to flag ungradable images or out-of-distribution (OOD) data, and (3) the model's ability to maintain high performance when data in the deployment setting is slightly or locally unaligned with the training distribution. We developed and validated approaches to address these robustness facets in radiology and ophthalmology applications. For radiology, we focused on the rapid triage of pulmonary coronavirus disease 2019 (COVID-19) using chest computed tomography (CT) scans, employing automated disease likelihood assessment and severity scoring. For ophthalmology, we devised and evaluated methods for increased robustness of DL systems for various retinal diseases and imaging modalities. These approaches were applied to systems for glaucoma screening using color fundus photographs (CFP), along with models for staging and biomarker quantification of age-related macular degeneration (AMD) from optical coherence tomography (OCT) scans.

CHAPTER 2 introduces CORADS-AI, a DL system for assessing pulmonary COVID-19 in chest CT scans using CO-RADS and CT severity scoring. This system comprises algorithms for lobe segmentation, CO-RADS classification, and a CT severity score for measuring the level of parenchymal involvement in each lobe. We evaluated the system under conditions close to the training distribution, using internal evaluation with eight independent observers, and external evaluation. We found high agreement between the system and the observers for the CO-RADS and CT severity scoring tasks on the internal test set and good generalizability to external data.

In **CHAPTER 3**, we present a more technical analysis of the CO-RADS grading model

from Chapter 2. In DL research, there is a predominant focus on developing novel architectural modifications, often at the expense of generalizability across different datasets. This trend neglects systematic comparisons among commonly used algorithm components. Therefore, we present a systematic analysis of commonly used approaches. This analysis includes the evaluation of both the two-dimensional and three-dimensional variants from seven commonly used classification algorithms. Moreover, we performed an ablation study for network weight initialization strategies, methods for including prior information, and ways of defining the model output format.

CHAPTER 4 introduces the Artificial Intelligence for RObust Glaucoma Screening (AIROGS) challenge, promoting the development of robust glaucoma screening algorithms capable of handling OOD and ungradable data. We presented a large dataset of around 113,000 CFPs from 60,000 patients and 500 screening sites. The challenge focused on both referable glaucoma classification, and ungradable image detection without the use of ungradable training data. We evaluated solutions from 14 teams and found they performed similarly to a set of 20 expert ophthalmologists and optometrists on our test set. Moreover, using three external datasets, we demonstrated highly robust performance on three other publicly available datasets. Popular methods amongst the winning teams included vision transformers, the inclusion of optic disc detection models in pre-processing, and domain knowledge-based OOD detection.

In **CHAPTER 5**, we propose Uncertainty-Based Instance eXclusion (UBIX), a method to enhance the robustness of DL classification models. This method provides robustness to locally OOD data, without the need for any additional training, either supervised or unsupervised, on such data. UBIX employs multiple-instance learning and uncertainty estimation to minimize the impact of corrupted instances during inference. While UBIX can potentially be applied to various classification tasks, our study focused on its use for staging AMD using OCT scans. We trained our models using data from a single scanner and then tested them on external datasets from various vendors, which included vendor-specific artifacts. Our results demonstrated that incorporating UBIX enhanced performance on these external datasets, surpassing the outcomes observed in models without UBIX.

In **CHAPTER 6**, we apply domain adaptation (DA) techniques for the increased robustness of segmentation models for biomarkers related to the intermediate stage of AMD in OCT. For segmentation, we used a model trained on data from one scanner. We developed a Cycle-Consistent Generative Adversarial Network (CycleGAN) to transform the visual characteristics of scans obtained from another scanner into the appearance of the scans used for training our segmentation model. This approach was evaluated on data from MACUSTAR, a large European prospective and multicenter study on the progression

of AMD. We showed improved segmentation performance when using DA in the processing pipeline and found that our model was more consistent across scanners than the human annotator.

CHAPTER 7 also addresses the issue of the large variability in OCT appearances between scanners. The variability in image resolution can impair the consistency in biomarker quantification both within and across studies. In this chapter, we aimed to provide a first step towards a standardized, high-fidelity OCT scan by using a super-resolution approach based on a 3D diffusion model. We proposed to condition this diffusion model on *en face* images, allowing for well-informed generative decisions. We conducted both qualitative and quantitative analyses to demonstrate the feasibility of employing this super-resolution method. Specifically, the technique was used to enhance OCT volumes with low B-scan density to volumes that are 8 times denser.

Nederlandse samenvatting

Deep learning (DL)-modellen voor medische beeldverwerking zijn het afgelopen decennium zeer populair geworden. Hoewel deze modellen in onderzoeksomgevingen doorgaans goed presteren, schieten ze vaak tekort in de echte wereld. Dit kan aanzienlijke negatieve gevolgen hebben op zowel het vertrouwen in deze algoritmen als hun gebruik in de kliniek. Een belangrijke oorzaak van deze verminderde prestaties is het ontbreken van intrinsieke mechanismen binnen de meeste systemen om robuustheid te bevorderen. In dit proefschrift richten we ons op drie aspecten van robuustheid: (1) het vermogen van het model om goed te presteren op data die zeer gelijkend of identiek is aan de verdeling van de trainingsdata, (2) het vermogen van het model om niet-beoordeelbare afbeeldingen of out-of-distribution (OOD) data te herkennen en (3) het vermogen van het model om goed te blijven presteren wanneer data in de implementatiecontext enigszins of lokaal verschilt van de verdeling van de trainingsdata. We hebben methoden ontwikkeld en gevalideerd om deze robuustheidsaspecten te onderzoeken voor toepassingen binnen de oogheelkunde en de radiologie. Binnen de radiologie richtten we ons op de snelle triage van pulmonaire coronavirus disease 2019 (COVID-19) door middel van computed tomography (CT) scans van de borstkas, waarbij we automatisch de waarschijnlijkheid en de ernst van de ziekte analyseerden. Binnen de oogheelkunde hebben we methoden ontwikkeld en geëvalueerd om de robuustheid van DL-systemen voor verschillende netvliesaandoeningen en beeldvormingsmodaliteiten te verbeteren. Deze methoden zijn toegepast op modellen voor zowel de screening van glaucoom door middel van kleurenfundusfoto's (KFF's), de herkenning van de verschillende stadia van leeftijdsgebonden maculadegeneratie (LMD) op basis van optische coherentietomografie (OCT) als de kwantificatie van LMD-gerelateerde biomarkers in OCT.

HOOFDSTUK 2 introduceert CORADS-AI, een DL-systeem voor de beoordeling van pulmonaire COVID-19 in CT-scans van de borstkas door middel van CO-RADS en een CT-ernstscore. Dit systeem bestaat uit algoritmen voor segmentatie van de longkwabben, CO-RADS-classificatie en de CT-ernstscore voor het meten van de hoeveelheid parenchy-

male betrokkenheid in iedere kwab. We evalueerden het systeem onder omstandigheden die vergelijkbaar waren met die tijdens het verkrijgen van de trainingsdata, met behulp van een interne evaluatie met acht onafhankelijke waarnemers en een externe evaluatie. De overeenstemming tussen de waarnemers voor de CO-RADS- en CT-ernstscore-taken was hoog in de interne dataset en het systeem generaliseerde goed naar de externe data.

In **HOOFDSTUK 3** presenteren we een technischere analyse van het CORADS-classificatiemodel van Hoofdstuk 2. In DL-onderzoek ligt de nadruk vooral op het ontwikkelen van nieuwe architectuuraanpassingen, wat vaak ten koste gaat van de generaliseerbaarheid over verschillende datasets. Deze tendens verwaarloost systematische vergelijkingen tussen al bestaande, veelgebruikte algoritmecomponenten. Daarom presenteren we een systematische analyse van veelgebruikte methoden. Deze analyse omvat de evaluatie van zowel tweedimensionale als driedimensionale varianten van zeven veelgebruikte classificatiealgoritmen. Daarnaast hebben we een ablatiestudie uitgevoerd naar initialisatiestrategieën van netwerkgewichten, methoden om a-priori-informatie te gebruiken en manieren om de vorm van de modeluitvoer te definiëren.

HOOFDSTUK 4 introduceert de Artificial Intelligence for RObust Glaucoma Screening (AIROGS) challenge, gericht op het promoten van de ontwikkeling van robuuste algoritmen voor glaucoomscreening die in staat zijn om OOD en niet-beoordeelbare data te herkennen. We presenteerden een grote dataset met ongeveer 113.000 KFF's afkomstig van 60.000 patiënten en 500 screeningslocaties. De challenge richtte zich op zowel de classificatie van verwijsbaar glaucoom als het herkennen van niet-beoordeelbare afbeeldingen, zonder gebruik te maken van niet-beoordeelbare trainingsdata. We evalueerden oplossingen van veertien teams en ontdekten dat ze op onze test set vergelijkbaar presteerden met een groep van twintig ervaren oogartsen en optometristen. Verder hebben we door middel van drie externe datasets laten zien dat de ontwikkelde algoritmen zeer robuust waren. Populaire methoden onder de winnende teams waren vision transformers, het gebruik van detectiemodellen voor de papil tijdens het voorbewerken van de data, en op domeinkennis gebaseerde OOD-detectie.

In **HOOFDSTUK 5** stellen we Uncertainty-Based Instance eXclusion (UBIX) voor, een methode om de robuustheid van DL-classificatiemodellen te vergroten. Deze methode biedt robuustheid tegen data die lokaal OOD is, zonder dat extra training, zowel supervised als unsupervised, met dergelijke data nodig is. UBIX maakt gebruik van multiple-instance learning en onzekerheidsschatting om de invloed van corrupte instances tijdens het toepassen van het model te minimaliseren. Hoewel UBIX in potentie kan worden toegepast op verschillende classificatieproblemen, richtte onze studie zich op het gebruik ervan voor het herkennen van LMD-stadia in OCT-scans. Onze modellen zijn getraind op

data van één scanner en vervolgens getest op externe datasets van verschillende leveranciers, welke artefacten bevatten die specifiek bij die leveranciers voorkwamen. Onze resultaten lieten zien dat het gebruik van UBIX de prestaties op deze externe datasets verbeterde, vergeleken met modellen zonder UBIX.

In **HOOFDSTUK 6** passen we domain adaptation (DA)-technieken toe om de robuustheid van segmentatiemodellen voor biomarkers gerelateerd aan het tussenstadium van LMD in OCT te verbeteren. We gebruikten een segmentatiemodel getraind met data van één scanner. Vervolgens hebben we een Cycle-Consistent Generative Adversarial Network (CycleGAN) ontwikkeld om de visuele kenmerken van scans van een andere scanner om te zetten naar die van de scans die gebruikt zijn voor het trainen van het segmentatiemodel. Deze aanpak is geëvalueerd op data van MACUSTAR, een grote Europese prospectieve en multicenterstudie naar de progressie van LMD. We hebben laten zien dat de segmentaties verbeterden bij gebruik van DA en dat het model consistenter was tussen scanners dan de menselijke analist.

HOOFDSTUK 7 richt zich ook op het probleem van de grote verschillen in OCT-beelden tussen scanners. De variabiliteit in beeldresolutie kan nadelige gevolgen hebben op de consistentie in biomarkerkwantificatie, zowel binnen als tussen studies. In dit hoofdstuk hebben we ons gericht op een eerste stap richting een gestandaardiseerde OCT-scan van hoge kwaliteit door middel van een superresolutiemethode gebaseerd op een driedimensionaal diffusiemodel. We stelden voor om dit diffusiemodel te conditioneren op *en-face*-beelden, wat het model dient te helpen met het maken van beter geïnformeerde generatieve beslissingen. We hebben zowel kwalitatieve als kwantitatieve analyses uitgevoerd om de haalbaarheid aan te tonen van een dergelijke superresolutiemethode. De ontwikkelde techniek is gebruikt om OCT-volumes met een lage B-scandichtheid te verbeteren tot volumes die acht keer dichter zijn.

Publications

Journal publications

C. de Vente*, L. H. Boulogne*, K. V. Venkadesh, C. Sital, N. Lessmann, C. Jacobs, C. I. Sánchez, B. van Ginneken, "Automated COVID-19 grading with convolutional neural networks in computed tomography scans: a systematic comparison," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 129–138, 2021.

C. de Vente, K. A. Vermeer, N. Jaccard, H. Wang, H. Sun, F. Khader, D. Truhn, T. Aimyshev, Y. Zhanibekuly, T.-D. Le, A. Galdran, M. A. González Ballester, G. Carneiro, D. R. G, H. P. S, D. Puthussery, H. Liu, Z. Yang, S. Kondo, S. Kasai, E. Wang, A. Durvasula, J. Heras, M. A. Zapata, T. Araújo, G. Aresta, H. Bogunović, M. Arikan, Y. C. Lee, H. B. Cho, Y. H. Choi, A. Qayyum, I. Razzak, B. van Ginneken, H. G. Lemij, C. I. Sánchez, "AIROGS: Artificial intelligence for robust glaucoma screening challenge," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 542–557, 2023.

C. de Vente, B. van Ginneken, C. B. Hoyng, C. C. Klaver, C. I. Sánchez, "Uncertainty-aware multiple-instance learning for reliable classification: Application to optical coherence tomography," *Medical Image Analysis*, vol. 97, pp. 103259, 2024.

C. de Vente, P. Valmaggia, C. B. Hoyng, F. G. Holz, M. M. Islam, C. C. W. Klaver, C. J. F. Boon, S. Schmitz-Valckenberg, MACUSTAR Consortium, A. Tufail, M. Saßmannshausen*, C. I. Sánchez*, "Generalizable Deep Learning for the Detection of Incomplete and Complete Retinal Pigment Epithelium and Outer Retinal Atrophy: A MACUSTAR Report," *Translational Vision Science & Technology*, vol. 13, no. 9, pp. 11, 2024.

N. Lessmann*, C. I. Sanchez*, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke, B. Geurts, H. A. Gietema, M. Groeneveld, L.

van Harten, N. Hendrix, W. Hendrix, H. J. Huisman, I. Isgum, C. Jacobs, R. Kluge, M. Kok, J. Krdzalic, B. Lassen-Schmidt, K. van Leeuwen, J. Meakin, M. Overkamp, T. van Rees Vellinga, E. M. van Rikxoort, R. Samperna, C. Schaefer-Prokop, S. Schalekamp, E. T. Scholten, C. Sital, L. Stöger, J. Teuwen, K. Vaidhya Venkadesh, **C. de Vente**, M. Vermaat, W. Xie, B. de Wilde, M. Prokop, B. van Ginneken, "Automated assessment of CO-RADS and chest CT severity scores in patients with suspected COVID-19 using artificial intelligence," *Radiology*, vol. 298, no. 1, pp. E18-E28, 2021.

R. Schwartz, H. Khalid, S. Liakopoulos, Y. Ouyang, **C. de Vente**, C. González-Gonzalo, A. Y. Lee, R. Guymer, E. Y. Chew, C. Egan, Z. Wu; H. Kumar, J. Farrington, P. L. Müller, C. I. Sánchez, A. Tufail, "A deep learning framework for the detection and quantification of reticular pseudodrusen and drusen on optical coherence tomography," *Translational Vision Science & Technology*, vol. 11, no. 12, pp. 3, 2022.

H. G. Lemij, **C. de Vente**, C. I. Sánchez, K. A. Vermeer, "Characteristics of a large, labeled dataset for the training of artificial intelligence for glaucoma screening with fundus photographs," *Ophthalmology Science*, vol. 3, no. 3, pp. 100300, 2023.

Conference papers

C. de Vente, M. M. Islam, P. Valmaggia, C. B. Hoyng, A. Tufail, C. I. Sánchez on behalf of the MACUSTAR Consortium, "Conditioning 3D Diffusion Models with 2D Images: Towards Standardized OCT Volumes through En Face-Informed Super-Resolution," *NeurIPS Workshop on GenAI for Health: Potential, Trust and Policy Compliance*, 2024.

M. M. Islam, **C. de Vente**, B. Liefers, C. Klaver, E. J. Bekkers, C. I. Sánchez, "Uncertainty-aware retinal layer segmentation in OCT through probabilistic signed distance functions," in *Medical Imaging with Deep Learning (MIDL)*, 2024.

Conference abstracts

C. de Vente, M. van Grinsven, S. De Zanet, A. Mosinska, R. Sznitman, C. C. W. Klaver, C. I. Sánchez, "Estimating uncertainty of deep neural networks for age-related macular degeneration grading using optical coherence tomography," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2020.

C. de Vente, C. González-Gonzalo, E. F. Thee, M. van Grinsven, C. C. W. Klaver, C. I. Sánchez, "Making AI transferable across OCT scanners from different vendors," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2021.

C. de Vente, K. Vermeer, N. Jaccard, H. G. Lemij, C. I. Sánchez, "AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge," in *Imaging and Morphometry Association for Glaucoma in Europe (IMAGE)*, 2022.

C. de Vente, A. Tufail, S. Schmitz-Valckenberg, M. Saßmannshausen, C. Hoyng, C. I. Sánchez on behalf of the MACUSTAR consortium, "OCT Super-Resolution for Data Standardization using AI: A MACUSTAR report," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2023.

A. Ardu, B. Liefers, **C. de Vente**, C. González-Gonzalo, C. C. W. Klaver, C.I. Sánchez, "Artificial Intelligence for the classification and quantification of reticular pseudodrusen in multimodal retinal images," in *European Society of Retina Specialists (EURETINA)*, 2020.

C. González-Gonzalo, E. F. Thee, B. Liefers, **C. de Vente**, C. C. W. Klaver, C. I. Sánchez, "Hierarchical curriculum learning for robust automated detection of low-prevalence retinal disease features: application to reticular pseudodrusen," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2021.

R. Schwartz, H. Khalid, S. Liakopoulos, Y. Ouyang, **C. de Vente**, C. González-Gonzalo, A.Y. Lee, C.A. Egan, C.I. Sánchez, A. Tufail, "A deep learning pipeline for the detection and quantification of drusen and reticular pseudodrusen on optical coherence tomography," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2022.

H. G. Lemij, **C. de Vente**, C. I. Sánchez, J. Cuadros, N. Jaccard, K. Vermeer, "Glaucomatous features in fundus photographs of eyes with 'Referable glaucoma' of a large population based labeled data set for training an Artificial Intelligence (AI) algorithm for glaucoma screening," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2022.

R. Schwartz, A. Olvera-Barrios, A. N. Warwick, H. Khalid, S. Phatak, M. Jhingan, **C. de Vente**, P. Valmaggia, C. I. Sánchez, C. A Egan, A. Tufail, "Increasing Stroke Risk Correlated with Higher Reticular Pseudodrusen Counts: Evidence from the UK Biobank," in *Association for Research in Vision and Ophthalmology (ARVO)*, 2024.

*equal contribution.

Author contributions

The main chapters in this thesis are based on publications, as referenced on the first page of each chapter. The contributions of the authors to these publications are listed below.

CHAPTER 2: Automated CO-RADS and chest CT severity scoring

Contribution	Authors
Guarantors of integrity of entire study	C.I.S., K.V.V., M.V., B.v.G.
Study concepts/study design or data acquisition or data analysis/interpretation	All authors
Manuscript drafting or manuscript revision for important intellectual content	All authors
Approval of final version of submitted manuscript	All authors
Agrees to ensure any questions related to the work are appropriately resolved	All authors
Literature research	N.L., C.I.S., M.B., C. Schaefer-Prokop, S.S., J.T., W.X., B.d.W., M.P., B.v.G.
Clinical studies	L.B., M.B., W.M.v.E., H.G., H.H., J.K., T.v.R.V., S.S., E.T.S., M.V., M.P., B.v.G.
Statistical analysis	N.L., C.I.S., R.S., C. Schaefer-Prokop, B.v.G.
Manuscript editing	N.L., C.I.S., L.B., L.H.B., M.B., J.P.C., T.D., W.M.v.E., H.G., N.H., W.H., I.I., C.J., J.K., B.L., K.v.L., J.M., T.v.R.V., E.M.v.R., R.S., C. Schaefer-Prokop, S.S., E.T.S., J.L.S., J.T., K.V.V., W.X., B.d.W., M.P., B.v.G

CHAPTER 3: A systematic comparison of COVID-19 grading models

Contribution†	Authors
Conceptualization, Methodology, Writing - review & editing	All authors
Software	C.d.V., L.H.B., K.V.V., C.S., N.L.
Validation, Formal analysis, Investigation, Writing - original draft	C.d.V., L.H.B., K.V.V., C.S.
Resources, Supervision, Project administration, Funding acquisition	C.J., C.I.S., B.v.G.
Data Curation	C.d.V., L.H.B., B.v.G.
Visualization	C.d.V., L.H.B.

CHAPTER 4: International challenge on robust glaucoma screening

Contribution†	Authors
Conceptualization	C.d.V., K.A.V., N.J., H.G.L., C.I.S.
Methodology, Writing - review & editing	All authors
Software	C.d.V., K.A.V.
Validation, Formal analysis, Investigation	C.d.V., K.A.V., H.G.L., C.I.S.
Resources	K.A.V., B.v.G., H.G.L., C.I.S.
Data curation	C.d.V., K.A.V., H.G.L.
Writing - original draft, Visualization	C.d.V.
Supervision, Funding acquisition	B.v.G., K.A.V., H.G.L., C.I.S.
Project administration	K.A.V., H.G.L., C.I.S.

CHAPTER 5: Reliable classification using Uncertainty-Based Instance eXclusion

Contribution†	Authors
Writing – review & editing	All authors
Writing – original draft, Visualization, Software	C.d.V.
Validation, Methodology, Investigation, Formal analysis, Conceptualization	C.d.V., C.I.S.
Data curation	C.d.V., C.B.H., C.C.W.K.
Supervision, Funding acquisition	B.v.G., C.I.S.
Resources, Project administration	C.I.S.

CHAPTER 6: Generalizable iRORA and cRORA detection models

Contribution†	Authors
Conceptualization	C.d.V., P.V., M.M.I., S.S.-V., A.T., M.S., C.I.S.
Methodology	C.d.V., M.M.I., A.T., M.S., C.I.S.
Software, Visualization	C.d.V.
Validation, Formal analysis, Investigation	C.d.V., M.S., C.I.S.
Resources	C.B.H., F.G.H., C.C.W.K., C.J.F.B., S.S.-V., M.C., A.T., C.I.S.
Data curation	C.d.V., C.B.H., F.G.H., C.C.W.K., C.J.F.B., S.S.-V., M.C., A.T., M.S., C.I.S.
Writing - original draft	C.d.V., M.S.
Writing - review & editing	All authors
Supervision	C.I.S.
Project administration	M.C., A.T., C.I.S.
Funding acquisition	F.G.H., S.S.-V., M.C., A.T., C.I.S.

CHAPTER 7: Standardizing OCT through *en face* conditioned 3D diffusion models

Contribution†	Authors
Conceptualization, Methodology	C.d.V., M.M.I., P.V., C.I.S.
Software, Investigation, Data curation, Writing - original draft, Visualization	C.d.V.
Validation, Formal analysis	C.d.V., C.I.S.
Writing - review & editing	All authors
Supervision, Project administration, Funding acquisition	C.I.S.
Resources	C.B.H., A.T., C.I.S.

† Following the Contributor Role Taxonomy (CRediT), see `https://credit.niso.org/`.

Contributed resources

We made several contributions related to the publications this thesis was based on in the form of publicly released source code, trained algorithms that can be run online using a web interface, and challenges that are still open for submission:

Source code

- The source code used to evaluate algorithms for the AIROGS challenge from Chapter 4:
  `https://github.com/qurAI-amsterdam/airogs-evaluation`
- The source code with an example algorithm to submit to the AIROGS challenge from Chapter 4:
  `https://github.com/qurAI-amsterdam/airogs-example-algorithm`
- The source code from a subset of the teams participating in the AIROGS challenge from Chapter 4:
  `https://airogs.grand-challenge.org/#final-leaderboard-at-isbi-2022`
- The source code for the UBIX algorithm presented in Chapter 5:
  `https://github.com/qurAI-amsterdam/ubix-for-reliable-classification`

Algorithms

- The trained CORADS-AI algorithm presented in Chapter 2:
  `https://grand-challenge.org/algorithms/corads-ai/`.

- The trained algorithms from the teams participating in the AIROGS challenge from Chapter 4:
  `https://airogs.grand-challenge.org/#final-leaderboard-at-isbi-2022`
- The trained UBIX algorithm presented in Chapter 5:
  `https://grand-challenge.org/algorithms/amd-classification-in-oct-with-ubix/`

Challenges

- A challenge to compare other solutions to our COVID-19 classification algorithm on the external public iCTCF dataset from Chapter 3:
  `https://covid19.grand-challenge.org/`
- The AIROGS challenge presented in Chapter 4:
  `https://airogs.grand-challenge.org/`

Acknowledgements

In this part of my thesis, I want to thank everyone who helped me complete my PhD. Most puzzle pieces to finish this project I never would have found or connected without the help of others. And then there were those whose essential role was cheering me on every time I managed to snap a new piece into place. Below, I would like to deeply thank the people who did at least one of those things.

I would like to start by expressing my profound gratitude to my daily supervisor and promotor, Clarisa. Thank you for being such a nice supervisor. Much of what I have learned during my PhD is thanks to you, especially when it comes to your technical and clinical knowledge, presenting research, and converting loose ideas and results into papers. I also owe numerous professional connections to your network, many of whom I will acknowledge later. I always left our progress meetings with new positivity, ideas, clarity, and motivation. During tougher times, whether facing technical challenges related to research or dealing with bureaucratic and project-related obstacles, your support made all the difference. Thank you for allowing me the freedom to explore while guiding me back on track when needed. I truly could not have wished for a better supervisor. And thank you for not being fed up with me yet and continuing to work with me after my PhD!

My gratitude extends to my promotor, Bram, for the opportunity to start my PhD in the unique and amazing research group you set up. Your feedback always hits the nail on the head. You can quickly pinpoint the remaining points for improvement, from the finest details to the bigger picture. I think your dedication to the field of medical image analysis is truly inspiring and I look forward to our new adventures in our future projects together!

I express my appreciation to the members of my PhD committee, prof. dr. Cees Snoek, prof. dr. Raphael Sznitman, prof. dr. Camiel Boon, prof. dr. Josien Pluim, prof. dr. Caroline Klaver, and prof. dr. Ivana Išgum, for reading and critically assessing my thesis.

Throughout my PhD journey, I've had the opportunity to work in several offices and labs, each made up of amazing people.

I would like to thank my teammates from the Eye Lab (previously known as the Retina

group): Cristina, Bart, Alessandro, Mohammad, and Roy. When I joined DIAG in 2019, I was very warmly welcomed by Cristina and Bart, a power couple that has done great work for the group. Bart, thank you for giving me my first introduction into the OCT world. Cristina, thank you for exploring our new workplace in Amsterdam together. This lab is always there but typically not so big – at one point I was the only PhD student there, right before another great colleague joined the Eye Lab, Mohammad. In some respects, we can be quite different: you're particularly enthusiastic about very theoretical topics, while I'm often a bit more practically oriented. However, that never stops us from having engaging and thought-provoking discussions. I'll never forget our great visits to London and Paris, including the conferences themselves and the nice dinners, but also that one time waiting in the middle of the night for the key in front of our AirBnB. And of course, thanks for agreeing to be my paranymph. Roy, even though we've only met in person once, it doesn't feel that way somehow. Thanks for our enjoyable video calls, your Obsidian tips and tricks, and everything else I've learned from you.

Thanks to everyone who has been a member of the qurAI group and the BMEP department during my PhD for the nice lunches together, our bouldering sessions, the great discussions we've had together, and your valuable feedback: Andrés, Ajey, Álvaro, Bob, Carolina, Caroline, Dimitrios, George, Gino, Hoel, Ioannis, Jin, Joaquim, Jörg, Juanjo, Laura, Louis, Lucas, Mahsa, Maria, Marinka, Matej, Michel, Navchetan, Nils, Piero, Riaan, Roel, Sanne, Sanyog, Shannon, Simone, Soufiane, Thomas, Vikram, Xiaojin, and Zhiwei. And Michel, thanks for being my paranymph as well!

Thank you to everyone at DIAG for the nice discussions, the lunches and borrels at the start of my PhD, and of course the outings and DIAG weekends: Anindo, Anne, Bram, Colin, Daan, Ecem, Erdi, Fazael, Francesco, Gabriel, Geert, Henkjan, James, Jasper, Kevin, Kicky, Matin, Nils, Patrick, Riccardo, Stephan, Thomas, Ward, Wouter, and Xie. Special thanks to the team of PhD students in the CORADS-AI classification project: Luuk, Cheryl, and Kiran. I had a great time working on this project together with you in such a collaborative way, something that can sometimes be missing in our usual projects. Kiran, thanks for your positivity and the good times we've had together, both during and outside of work. And, of course, thank you for involving me in your new entrepreneurial adventures. I'm really impressed with how much you have achieved since you arrived in the Netherlands and how you are now taking up a leading role. My gratitude further extends to everyone at the RSE team at DIAG for all the great supporting work you are doing to make our work as researchers a bit easier and more impactful.

I would also like to thank everyone I've worked with at the AMC Oogheelkunde department for the great collaborations and clinical insights you gave me: Camiel, Jonathan, Hind,

and Yasmine.

Thanks to everyone at the CIVIC AI lab who welcomed us so warmly when our group moved to UvA: Dimitris, Ilse, Katerina, Mayesha, Mirthe, Sara, and Tim. You were among our first colleagues we got to meet in Amsterdam, and your friendliness, our enjoyable lunches, and drinks together made the transition much better.

My gratitude further extends to everyone that participated in our EyeNED meetings on Thursdays for all your feedback and discussions: Caroline, Daniël, Eric, Haras, Harm, Jeroen, Jos, Jose, Karin, and Timo.

Thanks to all collaborators in the Macustar project, in particular the great colleagues that I've worked very closely with during this project: Adnan, Philippe, Luis, and Marlene.

I would like to deeply thank Hans and Koen from the Rotterdam Eye Hospital and Nicolas at Orbis for working together on the AIROGS challenge, which definitely would not have been possible without you. Also, thanks to all participating teams in the challenge for coming up with your great solutions, hopefully bringing glaucoma prevention a few steps closer using cost-effective screening.

Thanks to Pieter and Mitko for mentoring me during my MSc project, which helped a lot to prepare for my PhD. And of course, thanks to Pieter for the first step towards this PhD project by referring me to Clarisa.

Bedankt aan mijn familie en vrienden voor jullie steun en de nodige afleiding van mijn werk. In het bijzonder wil ik graag mijn ouders bedanken voor al jullie steun en liefde. Jullie begrijpen misschien niet altijd alle details van wat ik doe, maar toch heel erg bedankt voor jullie interesse in mijn werk. En dan bedoel ik in het bijzonder natuurlijk papa, die altijd heel geïnteresseerd is in of er nog ontwikkelingen zijn aan het front. En Ilse, Fleur, Thijs, en Olivia (mijn lieve zus, zusje (inclusief aanhang ;)) en nichtje), dankjulliewel voor die eerder genoemde afleiding van mijn werk. Met jullie is het altijd gezellig en ik zou niet weten wat ik zonder jullie zou moeten!

En dan als laatste, maar zeker niet *least*, Tamar. Ik ben superblij dat ik jou heb leren kennen. Ik denk dat ik niet overdrijf als ik zeg dat je het beste bent wat me is overkomen tijdens mijn PhD. Terwijl ik dit laatste stukje van mijn dankwoord schrijf, zijn we samen aan het reizen door het prachtige Indonesië. Dat is één van de mooie avonturen die we samen hebben meegemaakt en waarvan ik hoop dat er nog ontelbaar meer zullen volgen. Ik hou van jou.

