Discovery of Biomarker Candidates in Retinal OCT Images using Deep Learning

DOCTORAL THESIS

at the Medical University of Vienna for obtaining the academic degree

Doctor of Philosophy (PhD)

submitted by

Dipl.-Ing. Philipp Seeböck, BSc

Supervisor:
Assoc.-Prof. Dipl.-Ing. Dr. Georg Langs
Computational Imaging Research Lab,
Department of Biomedical Imaging and Image-guided Therapy
Medical University of Vienna

Vienna, 04/2019

Medical University of Vienna
A-1090 Wien ▪ Spitalgasse 23 ▪ Tel. +43-1-40160-0 ▪ www.meduniwien.ac.at

For Harald.

Abstract

BIOMARKERS constitute an essential building block of precision medicine, since they are important for diagnosis, treatment guidance and patient management. At the same time, there exists a lack of effective biomarkers in various diseases, highlighting the importance of biomarker discovery. Optical coherence tomography (OCT) is a non-invasive imaging modality that allows to assess morphological conditions and changes of the retina, and is currently one of the most important diagnostic modalities in ophthalmology. In particular, age-related macular degeneration (AMD) is one of the leading causes of blindness in the world. While its clinical signs are observable in OCTs, the underlying pathogenic mechanisms are not yet fully understood, meaning that there might be relevant structures that have not been discovered so far. In this context, automated medical image analysis approaches offer the potential of efficiently exploring imaging data to detect and evaluate biomarkers. Particularly deep learning is a powerful branch of state-of-the-art data-driven machine learning techniques, which are capable of learning complex relationships directly from data.

In this thesis, the main purpose is to develop and evaluate novel deep learning techniques for automated identification of new biomarker candidates in retinal OCT images, without the use of manual labels. First we propose an unsupervised deep learning method that is trained on unlabeled data to learn healthy anatomical appearance for detection and categorization of anomalies, which form biomarker candidates. We demonstrate that the identified marker candidates are stable, show predictive value in the task of detecting disease and align with our current understanding of disease course. The second developed method exploits a novel way to improve anomaly detection in retinal OCT images, using Bayesian deep learning. Information about healthy anatomical appearance is jointly used with epistemic uncertainty estimates to detect deviations from normal, achieving results that clearly outperform other state-of-the art methods. The third method is trained in an unsupervised way on large amounts of unlabeled data to learn disease specific features from OCTs. The model both captures local characteristics of the retina and learns low-dimensional global representations of whole OCT volumes. We demonstrate that the learned features correlate well with already known biomarkers, as well as features that had not been considered yet in clinical practice, i.e. features that form new biomarker candidates. Additionally, the results show that the learned features correlate better with visual function than established makers.

The proposed methods were trained and evaluated on OCT imaging data of the human retina, particularly in patients suffering from AMD. We demonstrate that the proposed techniques are effective for identifying disease marker candidates in retinal OCT images.

Kurzfassung

Biomarker sind ein wesentlicher Bestandteil der Präzisionsmedizin, da sie sowohl für Diagnose und Therapie als auch für Patientenmanagement wichtig sind. Gleichzeitig fehlt es bei vielen Krankheiten an effektiven Biomarkern, was die Bedeutsamkeit der Erforschung und Entdeckung von Biomarkern unterstreicht. Optische Kohärenztomografie (OCT) ist eine nicht-invasive Bildgebungsmodalität, mit der morphologische Zustände und Veränderungen der Netzhaut beurteilt werden können, und ist derzeit eine der wichtigsten diagnostische Modalitäten in der Ophthalmologie. Insbesondere ist altersbedingte Makuladegeneration (AMD) eine der weltweit häufigsten Ursachen für Erblindung. Während die klinischen Symptome in OCT Bildern sichtbar werden, sind die zugrunde liegenden pathogenen Mechanismen noch nicht vollständig erforscht. Dies bedeutet, dass relevante Strukturen im Bild vorhanden sein könnten, deren klinischer Wert bis jetzt noch nicht erkannt wurde. In diesem Zusammenhang bieten automatisierte Methoden der medizinischen Bildanalyse die Möglichkeit Bilddaten effizient zu erforschen, Biomarker zu erkennen und zu analysieren. Insbesondere *Deep Learning* ist ein leistungsfähiger Zweig von State of the Art datengetriebenen *Machine Learning* Methoden, die in der Lage sind komplexe Beziehungen direkt aus den Daten zu lernen.

Das Hauptziel dieser Doktorarbeit besteht in der Entwicklung und Evaluierung neuartiger Deep Learning Methoden zur automatisierten Identifizierung neuer Biomarker-Kandidaten in OCT Bildern der menschlichen Netzhaut. Zunächst stellen wir ein *unüberwachtes Lernverfahren* vor, welches auf Daten ohne manuelle Annotierungen trainiert wird. Dabei wird das normale anatomische Erscheinungsbild gelernt, um dann Anomalien erkennen und kategorisieren zu können, welche Biomarker-Kandidaten darstellen. Wir zeigen, dass die identifizierten Markerkandidaten stabil erfasst werden, einen prädiktiven Wert bei der Erkennung von Krankheiten aufweisen und mit unserem derzeitigen Verständnis des Krankheitsverlaufs übereinstimmen. Die zweite entwickelte Methode nutzt einen neuartigen Ansatz, um die Erkennung von Anomalien in retinalen OCT-Bildern mithilfe von *Bayesian Deep Learning* zu verbessern. Informationen über ein gesundes anatomisches Erscheinungsbild werden gemeinsam mit Schätzungen der bayesschen Unsicherheit verwendet, um Abweichungen vom Normalzustand (Anomalien) zu erkennen. Dabei zeigen die Ergebnisse, dass die entwickelte Methode besser als andere State of the Art Ansätze in Bezug auf Anomaliedetektion funktioniert. Die dritte Methode wird auf großen Datenmengen ohne manuelle Annotierungen trainiert (*unüberwachtes Lernen*), um krankheitsspezifische Merkmale aus OCT-Bildern zu lernen. Das Modell erfasst lokale Merkmale der Netzhaut und ermöglicht gleichzeitig eine niedrigdimensionale Darstellung des gesamten OCT-Volumens. Es wird gezeigt, dass die gelernten Merkmale gut mit bereits bekannten Biomarkern korrelieren, welche aktuell in der klinischen Praxis verwendet werden. Gleichzeitig werden auch neue Merkmale gelernt, die in der klinischen Praxis bis jetzt noch nicht berücksichtigt wurden (d.H. Merkmale, die neue Biomarker-Kandidaten bilden). Darüber hinaus zeigen die Ergebnisse, dass die erlernten Marker-Kandidaten besser mit der Sehleistung korrelieren als etablierte Marker.

Die in dieser Doktorarbeit entwickelten Methoden wurden an OCT-Bilddaten der menschlichen Netzhaut trainiert und ausgewertet. Hierbei vorrangig bei Patienten, die an AMD leiden. In diesem Zusammenhang zeigen wir, dass die präsentierten Methoden in der Lage sind neue Biomarker-Kandidaten in retinalen OCT-Bildern zu entdecken.

Acknowledgments

Writing the acknowledgments is probably the most important part of a doctoral thesis, since it would not have been possible to reach this point without the support and help of other people. There are moments when you have to solve problems you have never thought of before, moments that require a lot of patience, energy and perseverance, or moments where you face challenges that seem unsolvable at a first glance. In all of these situations the encouragement of other people is crucial to overcome the difficulties and to find a way through the maze.

First of all I would like to thank my supervisor Georg Langs for advising me during the last years. Georg, I have learned a lot from you both from a technical and human perspective. I am grateful for every minute you spend for me, since each single one helped me to develop and improve this thesis and my work in the field of research in general. You are exactly the supervisor that everybody wants to have - thank you.

I would also like to thank the members of my thesis committee Sebastian Waldstein and Wolfgang Birkfellner for their valuable comments and fruitful discussions that helped me to improve this thesis. Moreover, I want to thank you, Sebastian, also for your support in your role as project leader of OPTIMA. I always appreciate your advise. In this context, I am especially grateful for having the opportunity to work in this interdisciplinary group. It was not only a great experience from a professional perspective, but at least as much from the human point of view. I would like to thank Bianca, the second project leader of OPTIMA, in particular, because you have always taken time for me. This is not self-evident and I will never forget how you supported me even in difficult times.

I also want to express my thanks to all my colleagues from CIR and OPTIMA. Everyone contributed in a particular way to my research, and some of you even became friends: Alessio, Amir, Antoine, Christoph, Christian, David, Ernst, Hrvoje, Jeanny, Jing, Johannes, Karl-Heinz, Luca, Martin, Markus H., Markus K., Mustafa, Nacho, René, Rona, Roxanne, Sophie, Thomas A., Thomas S. and Wolf-Dieter. Furthermore, I would like to thank all the awesome people that I met at scientific conferences, whom I could share inspiring thoughts and experiences with.

I also want to thank my amazing friends. Life so much better when you are around.

Big hugs to my family, of course. Thank you mum and dad for supporting me whatever I do and being the best parents I can imagine. To you, Lisa, for being the best sister in the world. To my grandparents, godparents and all the other family members for always supporting me when I need it.

From the bottom of my heart, I want to thank my wonderful wife Désirée, the love of my life. Thank you for being who you are.

Contents

List of Figures

List of Tables

List of Acronyms

AE	autoencoder
AI	artificial intelligence
AMD	age-related macular degeneration
ANN	artificial neural network
AuC	area under ROC curve
BGD	batch gradient descent
BM	Bruch's membrane
CAE	convolutional autoencoder
CAM	class activation mapping
CE	cross entropy
CNN	convolutional neural network
CRT	central retinal thickness
CT	computed tomography
DBN	deep belief network
DME	diabetic macular edema
DNN	deep neural network
DR	diabetic retinopathy
ELU	exponential linear unit
FA	fluorescein angiography
FPR	false positive rate
GA	geographic atrophy
GAN	generative adversarial network
GC	ganglion cell layer
GMM	Gaussian mixture model
GPU	graphics processing unit
INL	inner nuclear layer
IPL	inner plexiform layer
IS	inner segment
KL	Kullback-Leibler
LLVA	low luminance visual acuity
LSTM	long short term memory
MAE	mean absolute error
MAP	maximum a posteriori probability
MBGD	mini-batch gradient descent
MC	Monte Carlo
MRT	magnetic resonance imaging
MSE	mean squared error
NN	neural network
OC-NN	one class neural network
OCT	optical coherence tomography
ONL	outer nuclear layer
OPL	outer plexiform layer
OS	outer segment
PCA	principal component analysis
PEDF	pigment epithelium derived factor
PET	positron emission tomography
PR	precision-recall
PRN	pro re nata

RBM	restricted Boltzmann machine
ReLU	rectified linear unit
RF	random forest
ROC	receiver operating characteristic
RPE	retinal pigment epithelium
RVO	retinal vein occlusion
SD-OCT	spectral domain OCT
SGD	stochastic gradient descent
SRF	subretinal fluid
SVM	support vector machine
TNR	true negative rate
TPR	true positive rate
VA	visual acuity
VAE	variational autoencoder
VAR	visual acuity rating
VEGF	vascular endothelial growth factor
WHO	World Health Organization

CHAPTER 1

Introduction

"Most people fail in life
not because they aim too high and miss,
but because they aim too low and hit."

– Les Brown

In this thesis, we present novel conceptual and methodological contributions in the field of biomarker discovery. In particular, the proposed approaches provide different strategies to identify marker candidates. They are based on machine learning techniques and are evaluated in the context of retinal images, namely optical coherence tomography (OCT) scans. First, Section 1.1 provides a compact description of the motivation. Second, we give a general introduction into the topic of medical image analysis and machine learning in Section 1.2. Third, the problem definition and aims of this thesis are presented in Section 1.3. Finally, the contributions of our work are summarized in Section 1.4 and a thesis outline is provided.

1.1 Motivation

A systematic review of the World Health Organization (WHO) revealed that 285 million people are affected by visual impairment worldwide, with 246 million suffering from low vision and 39 million from blindness (Pascolini and Mariotti, 2012). Due to the global phenomenon of natural aging and the number of humans 50 years or older is 65% for all visual impaired and 82% for all blind people, the prevalence of these cases is expected to grow in the future (Pascolini and Mariotti, 2012). In Europe alone, the WHO reported in 2010 that nearly 3 million were blind and over 28 million were affected in total by visual impairment (Pascolini and Mariotti, 2012, Prokofyeva and Zrenner, 2012).

Several meta-analyses and large-scale population-based studies have shown that among the leading causes of blindness in Europe, AMD is the most frequent (26%), followed by

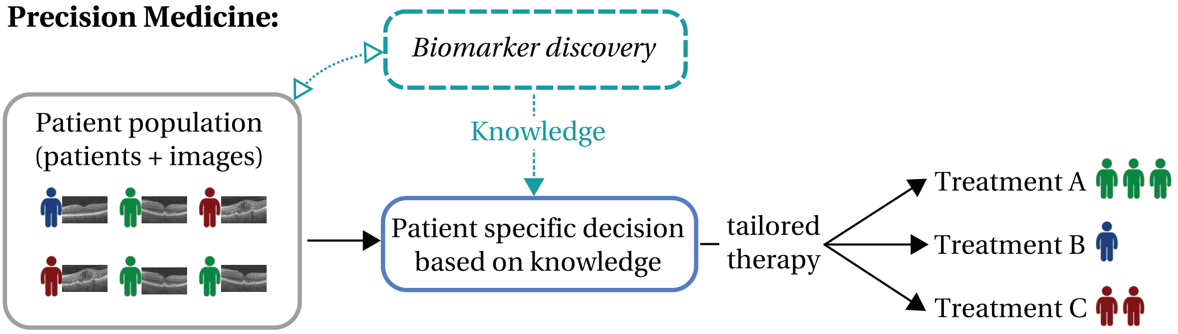


Figure 1.1: Precision medicine workflow and the role of biomarker discovery. The basic idea of precision medicine is to perform treatments optimized for each patient, based on individual characteristics and biomarker manifestations. The identification of hitherto unknown imaging biomarkers improves precision medicine and therefore individual patient care, in particular diagnosis and clinical decision processes. The aim of this thesis is to develop and explore innovative methods based on machine learning to detect new biomarker candidates in imaging data. This supports a critical shift in medicine, where imaging data and medical image analysis are jointly used for proactive hypothesis development.

glaucoma (20.5%) and diabetic retinopathy (DR) (8.9%) (Prokofyeva and Zrenner, 2012). With 25-30 million affected people, AMD is the most common cause of severe vision loss worldwide, showing a prevalence of 9% (Prokofyeva and Zrenner, 2012, Wong et al., 2014).

Since early stages of AMD usually start without symptoms, it is prone to late diagnosis at a level where pathological processes have already altered the healthy structure of the retina. Moreover, traditional techniques for diagnosing AMD such as the visual acuity test (Section 2.1.2) or the Amsler grid test (Fine et al., 2000) detect AMD at a stage where vision has already been affected by the disease. This means that people already suffer from a vision loss at the timepoint of initial treatment. In addition, 44% of all people show some degree of vision loss between the referral assessment and the beginning of treatment, where this timespan is around 28 days in Europe (Klein et al., 2008). This is of particular relevance, since vision loss is often not retrievable (Schmidt-Erfurth et al., 2018b).

OCT (Huang et al., 1991) provides a detailed three-dimensional high-resolution image of the retina and allows to inspect its condition at a μm resolution. It is a non-invasive imaging technique and currently the most important diagnostic modality in ophthalmology. It is widely used in the clinic, with 30 million OCT acquisitions conducted annually, or one scan carried out every second worldwide. These numbers are on par with other imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRT) or positron emission tomography (PET) (Fujimoto and Swanson, 2016). Among others, AMD manifests in changes that can be observed within OCTs. Hence, the scans are analyzed by physicians to perform diagnosis, determine treatment or infer other clinical decisions (Fujimoto and Swanson, 2016). However, limited predictive capability is currently inherent in known clinical signs that are visible (Schmidt-Erfurth et al., 2018b). Moreover, the underlying pathogenic mechanisms in

AMD are not yet fully understood (Schmidt-Erfurth and Waldstein, 2016), which leads to the conclusion that there might be unknown patterns or structures that have not been discovered so far. Another key factor is that researchers are overwhelmed by the massive amount of data, given the large number of OCTs in connection with millions of pixels per volume. As a consequence, manual inspection and analysis of these scans is practically impossible on a large-scale basis. In addition, diagnostics and clinical decisions may vary between physicians due to differences in professional experience, image quality, workload, time restrictions and the lack of consensus regarding the relevant imaging biomarkers. These markers are not only needed to enable an efficient management of the leading diseases such as diabetic retinopathy or age-related macular degeneration, but also to allow diagnosis and consequently treatment at the earliest possible stage, optimized for individual patients (Figure 1.1). Early diagnosis and an individualized treatment can be critical with respect to the prevention of potentially irreversible loss of function such as central vision, the success of therapy, reducing the overall burden of patients and lowering the financial load on the clinical health care system. In other words, the discovery of expressive imaging biomarkers is an essential building block towards precision medicine (Wang et al., 2017).

In this context, medical image analysis and machine learning offer the potential of automated exploration of large-scale imaging data to identify, analyze and evaluate new biomarker candidates, as discussed in the following (Section 1.2).

1.2 Medical Image Analysis

In the field of medical image analysis, medical problems are analyzed based on biomedical image data and digital image analysis. The field of medical image analysis emerged in the early 90s as its own separate new discipline (Wells III, 2016). This evolution was driven by the unique problem setting and characteristic of medical imaging in this area of study. The high-dimensional nature of imaging data, the statistic variability of normal and abnormal recordings, the heterogeneity of physical and physiological properties of the human body which are measured, the various types of image information that are acquired, the nonrigid characteristic of object motion and deformation as well as the three-dimensional property of image data are some examples for this uniqueness (Duncan and Ayache, 2000). These special characteristics together with additional potential requirements such as using image analysis as part of the clinical workflow constrains both the analysis and the selection of applicable algorithms (Toennies, 2017).

In this context, the application of machine learning methods allows to analyze large amounts of image data in an automated way (Duncan and Ayache, 2000, Lambin et al., 2012). Machine learning is defined as the field of algorithms that involve autonomous learning of a model based on data, to progressively improve the performance on a given task. A more detailed explanation of machine learning concepts is provided in Section 3.1. For instance, machine learning methods enable the generation of results that are objective and reproducible, allow to increase diagnosis accuracy or to visualize complex correlations and patterns (Doi, 2007, Langs et al., 2011, Wang et al., 2010). Furthermore, morphological and functional properties as well as relationships are discovered and analyzed (Achterberg et al., 2014, Vogl et al., 2017b).

Recently, a specific subgroup of machine learning techniques has shown impressive results in various medical tasks: *deep learning*. For instance, Kooi et al. (2017) proposed the detection of mammographic lesions using large scale deep learning, performing on par with certified screening radiologists on a patch level. Grewal et al. (2018) proposed a deep learning based approach to conduct hemorrhage detection in CT scans, approaching the detection accuracy of radiologists. Dermatologist-level classification accuracy of skin cancer was achieved by a deep neural network in Esteva et al. (2017), which was trained on 129,450 labeled images. Rajpurkar et al. (2017) developed a deep learning algorithm for pneumonia detection in chest x-ray images that achieved performance comparable to radiologists.

In general, medical imaging and its analysis can help to identify sub-groups of patients with varying risk profiles, progression paths and treatment responses. This forms a cornerstone of precision medicine, targeting individualized treatment for each patient in order to obtain the best possible treatment response (European Society of Radiology, 2015, Wang et al., 2017). As illustrated in Figure 1.1, identifying expressive biomarkers is essential to distinguish sub-groups with differing treatment response, and is a primary challenge in precision medicine (European Society of Radiology, 2015, Wang et al., 2017).

1.3 Problem definition and thesis aims

Although impressive results have been achieved in the field of medical image analysis (Section 1.2), the overall progress in medical image analysis has been slower (Wells III, 2016). Besides the specific requirements and the unique problem setting in medical imaging, the lack of large-scale labeled data in many medical image analysis, which are available in computer vision on natural images poses another challenge (Wells III, 2016). The requirement of labeled training data can be a crucial limitation for multiple reasons. They can be costly or even unfeasible to obtain in some clinical settings, may suffer from inter- and intra-grader variability (Asman and Landman, 2011), or restrict the exploration of potential biomarkers to a pre-defined set of marker categories. In this thesis, we explore the potential of identifying biomarkers without the use of manual labels. Specifically, the thesis aims are as follows:

Aim 1 - Anomalies as a means to focus on disease effects. Develop a method to segment biomarker candidates on pixel level in an unsupervised way, omitting the need for manual annotations to train a model. The method should distinguish normal from abnormal image regions, assuming that the detected anomalies represent biomarker candidates capturing relevant disease characteristics (Chapter 5).

Aim 2 - Anatomical knowledge to enhance anomaly detection. Exploit knowledge about normal anatomy and its variability in addition to imaging data to improve anomaly detection (Chapter 6).

Aim 3 - Representing different levels of detail in imaging data. Represent both phenotypical and disease related image characteristics in a patient population to exploit it for biomarker discovery. This compact representation of imaging data should jointly capture disease effects and normal anatomical variability on different levels of detail to aid hypothesis generation (Chapter 7).

A main goal of this thesis is the exploration of different strategies, methods and models to identify biomarker candidates in medical images. Specifically, we develop deep learning techniques in the context of retinal OCT images that can be trained without manually annotated labels. Deep learning (Section 3.2) is a specific branch of machine learning capable of automatically learning feature representations from data, instead of relying on hand-crafted features for a given task. In this thesis, we develop and present three deep learning methods that tackle the aims listed above.

1.4 Contribution and thesis outline

The thesis is divided into 8 chapters. It begins with a general overview, including the motivation, an introduction to medical image analysis and scope of the thesis. In Chapter 2 we provide the clinical background of this work. The anatomy and physiology of the human eye is discussed, with particular focus on the retina and its anatomical components. We then describe techniques which are currently available to measure visual function. Furthermore, insights about retinal diseases and its progression pathways are provided, with special emphasis on age-related macular degeneration (AMD). State-of-the-art treatments are covered as well, including potentials and limitations of actual disease handling. We also provide a description of the imaging technique OCT, underlying its value for diagnosis and patient management in the field of ophthalmology. In the last part of Chapter 2, an introduction to clinical trial endpoints with focus on retinal studies is given.

Chapter 3 presents the methodological background of this thesis. First, machine learning fundamentals are presented. Training principles of machine learning models and involved components are described. Varying training paradigms such as supervised, weakly-supervised or unsupervised learning are introduced. Secondly, we give an overview about deep learning principles, since all methods proposed in this thesis are based on deep learning. To emphasize the difference to conventional machine learning approaches, basic building blocks, training and optimization dogmas, the principle of uncertainty estimations as well as more detailed descriptions of essential architectures of deep learning such as convolutional neural networks (CNNs) are provided. Finally, statistical evaluation measures that are commonly used in medical image analysis are discussed, forming the foundation with respect to the evaluation of methods conducted in this doctoral thesis.

The scope of Chapter 4 is to provide an overview about different strategies regarding image biomarker discovery. Advantages, potentials as well as limitations are discussed for each category.

The contribution of this thesis is structured around three manuscripts. The first one is presented in Chapter 5: *"Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data"*. An unsupervised deep learning method is proposed to identify marker candidates in imaging data. In a first step, anomaly detection is performed. In spectral-domain OCT images, anomalous candidates are separated from normal tissue based on features learned by an unsupervised deep auto-encoder on healthy samples, modeling the normal appearance distribution with a One-Class support vector machine (SVM). In a second step, stable categories of frequently occurring anomalies are identified using clustering, and their link to disease is evaluated. Results also demonstrate that these marker candidates show some predictive value in the task of detecting disease. While part of the categories could be mapped to known structures by retinal experts in a qualitative evaluation, others remain as novel data driven marker candidates.

The second manuscript *"Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT"* is presented in Chapter 6. Here, we propose a novel approach for anomaly detection exploiting uncertainty estimates of a Bayesian deep learning model. Weak labels (Section 3.1.3) are used to train a model for segmenting the anatomy of healthy subjects. Based on the assumption that uncertainties correlate with areas which are not present in the healthy training set, anomalies are detected. Results show that this technique outperforms previous state-of-the-art methods in anomaly detection in OCTs.

Chapter 7 presents the third manuscript *"A paradigm shift in retinal biomarker identification by unsupervised deep learning"*. Therein a two-level unsupervised deep learning approach is trained on a large-scale dataset of retinal OCT images. While the first level learns a local representation of morphology, the second level forms a global low-dimensional representation of the whole OCT volume. Results demonstrate that some of the identified features correlate well with already known biomarkers traditionally used in clinical practice, while others form new biomarker candidates that had not been considered yet in clinical practice. Furthermore, results show that the learned features correlate better with visual function than the established makers.

Finally, we provide a discussion and conclusion in Chapter 8, summarizing our methods and main results, relating the three presented methods with each other and pointing out future research lines that can be derived from our contributions.

CHAPTER 2

Clinical Background and Retinal Imaging

"No one can lie,
no one can hide anything,
when he looks directly into someone's eyes."

– Paulo Coelho

In this Chapter, the clinical background of our work is presented. We provide a brief explanation of the anatomy of the eye with focus on the retina (Section 2.1.1), an overview of techniques to measure visual function (Section 2.1.2), an introduction to specific retinal diseases (Section 2.1.3) and an overview of current treatment options (Section 2.1.4). Since the focus of this thesis lies on the discovery of imaging biomarkers in OCT scans, a description of OCT acquisition principles is provided in Section 2.2. Finally, an introduction to clinical trial endpoints with focus on retinal studies is given in Section 2.3, which provides an overview on the interaction of biomarker discovery and treatment of patients, study design as well as drug development in the context of retinal diseases.

2.1 Clinical Background

An introduction to the anatomy of the eye delivers the necessary background knowledge to understand the principles and characteristics of retinal diseases and relevant function measures, which are used to assess the functional impact of retinal diseases. Moreover, an overview and discussion of treatment options and its challenges provides the basis to understand the importance of biomarker discovery in the context of retinal diseases.

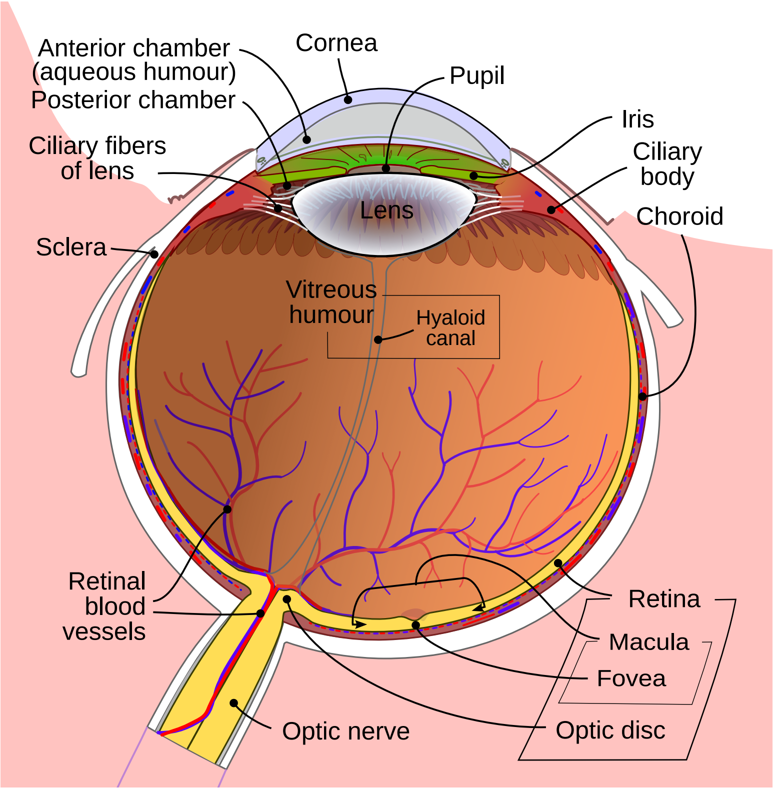


Figure 2.1: Schematic overview of the anatomy of the human eye. Source: Commons (2018b).

2.1.1 The Anatomy of the Eye

Along the various senses in humans, vision is among the most important (Rajamanickam, 2007). The complex system of visual sense involves two organs, the eyes and the brain. A schematic illustration of the eye is depicted in Figure 2.1. It is a spherical structure that is composed of two main parts, the anterior and posterior segments. The approaching light of the environment is focused by the anterior segment of the eye, while the posterior segment converts the light into electrical impulses that are sent to the brain (Savino and Danesh-Meyer, 2012).

The anterior segment is visible from outside and is composed of the cornea, the iris, the lens and the ciliary body (upper half of Figure 2.1) (Kaufman et al., 2011). The cornea is the front transparent surface of the eye and is the first structure that is hit by incoming light. Its main function is to refract the light towards the lens of the eye. The iris is a muscle tissue that surrounds the pupil, where the color of the iris is explained by the pigment cells present in these muscles. The pupil is an opening that is seen as a black circle, its diameter controlled by contraction and dilation of separate muscles. This determines the amount of light that enters the eye, allowing more when its darker or less in case of strong illumination (Kaufman et al., 2011). The lens has a biconvex form and focuses the light onto the retina. Its shape can be changed by contraction of the ciliary body, a muscle that is connected to the lens by the ciliary fibers. In addition to this function regarding accommodation, the cells of the ciliary body also secrete the fluid that fills both the anterior and posterior chamber of the eye. This fluid also known as aqueous humor provides nutrients to the ocular structures and outflows through the trabecular meshwork, a connective tissue. The intraocular pressure is determined by the amount of aqueous humor produced and drained off (Kaufman et al., 2011).

The posterior segment is surrounded by fat tissue that has a protective function, while a set of extrinsic extraocular muscles allows to move the eye. The posterior segment is composed of the vitreous and three layers enclosing it: the sclera, the choroid and the retina (Kaufman et al., 2011). The vitreous chamber contains a jelly-like material called vitreous humor that helps to maintain the spherical shape of the eye (Purves et al., 2001). The outermost layer surrounding the vitreous is the sclera, a white fibrous tunic that gives the eye its structural strength. Together with the cornea it makes up the outer tunic of the eye. The choroid is a vascular layer, supporting the eye with blood and oxygen using a network of thin capillaries (Kaufman et al., 2011). It is located between the sclera and the innermost layer, the retina.

The retina is a thin multi-layered sensory membrane that transforms the incoming light into nerve impulses which are then sent to the brain. Topographically, the main structures of the retina are the *optic disc*, the *macula* and the *fovea*. The optic disc is the location where the retinal blood vessels and the nerve fibers enter the eye. This is also known as blind spot, since this area does not contain any photoreceptors. The retinal vessels are responsible for approximately 20% of the blood needed by the retina, and provide nutrition from the top (Henkind et al., 1979, Hildebrand and Fielder, 2011). The remaining part of the blood is supplied from the bottom by the choroid. The macula is a yellow region of about 4.5-6mm in diameter in adults and accounts for almost all photopic vision (Provis et al., 2005). The central part of the macula is called the fovea, representing the region of greatest visual acuity (Oyster, 1999). This is explained by the fact that the fovea is characterized by the highest density of photoreceptors in the retina. Moreover, due to the high density of receptors no retinal capillaries are located in the central 0.5mm, meaning that this central region depends solely on the blood supply from the choroid. The fovea is about 0.35mm in diameter and is characterized by a depression in the retina (Figure 2.6, on the left, central depression of the ”normal retina”).

The cellular organization of the retina is depicted in Figure 2.2, showing its layered structure. The light enters from the vitreous, passes the ganglion cells and the interneurons before activating the photoreceptors in the outermost part of the retina near the choroid. Below the photoreceptors, the retinal pigment epithelium (RPE) acts as a blood-retina barrier, representing a selective barrier between the choriocapillaris and the photoreceptors. It is one of the most metabolically active tissues of the body and has a crucial supportive function for the photoreceptors. Bruch’s membrane (BM) separates the RPE from the choroid and acts as an molecular sieve: e.g. oxygen, nutrients or metabolic waste products are partly regulated by the BM (Hildebrand and Fielder, 2011, Kaufman et al., 2011).

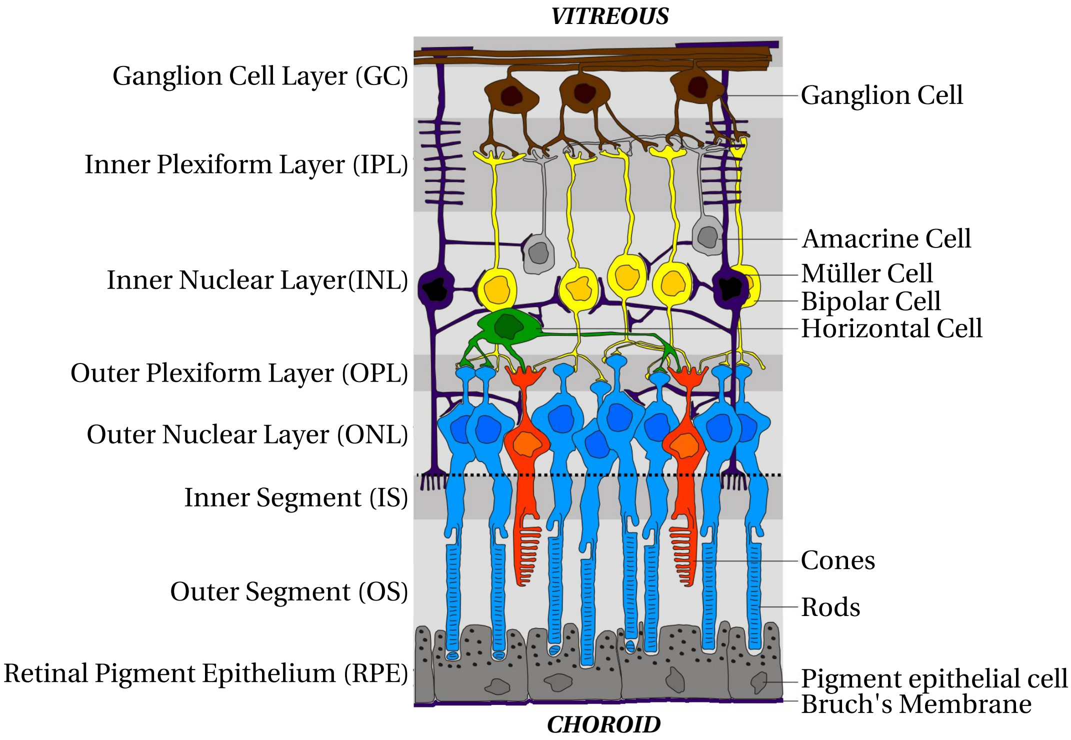


Figure 2.2: Schematic illustration of the layered structure of the human retina. Modified from source: Commons (2018a).

There are two types of photoreceptor cells: cones and rods. The human retina contains around 4-5 million cones and 77-107 million rods (Hildebrand and Fielder, 2011). Both types of photoreceptor cells have a long and elongated shape to maximize light sensitivity (Curcio et al., 1990). They consist of an outer segment, inner segment, a nucleus, an inner fiber and a synaptic terminal, which provide the basis for the layer names depicted in Figure 2.2 (outer segment (OS), inner segment (IS), outer nuclear layer (ONL), outer plexiform layer (OPL)). The photon-sensitive pigment that produces a biochemical signal is located in the outer segment, while the metabolic requirements of the photoreceptors are met by the inner segment. After transposing the biochemical into an electric signal, it is transferred via the inner fiber to the synaptic terminal located in the OPL. The electric signals of the photoreceptors are transmitted to the ganglion cells through the inner nuclear layer (INL) and the inner plexiform layer (IPL). The interneuron cells (bipolar, horizontal and amacrine) in these layers (INL, IPL) process the signals, integrating and regulating certain patterns before sending the resulting activation to the ganglion cells (Hildebrand and Fielder, 2011). Müller cells compose the main part of retinal glial cells in the INL, that serve as support for the interneuron cells. Finally, the visual information from the retina is transmitted by the ganglion cells through the optic nerve to the brain (Hildebrand and Fielder, 2011, Vilensky et al., 2015).

The first type of photoreceptor cells, the cones, enable to distinguish colors and are responsible for the fine-detailed view. They are mainly located in the fovea, the center of the macula, whereas rods are predominate in the peripheral part of the retina. Cones need a higher amount of illumination to get activated, meaning that they are crucial for vision under bright light conditions. On the other hand, rods are much more sensitive to light but are not able to distinguish different wavelengths. Since they register the intensity of the incoming light with high sensitivity and are not located in the fovea, rods are relevant for peripheral and scotopic vision (vision under dark conditions) (Hildebrand and Fielder, 2011).

In summary, the incoming light is refracted and focused onto the retina by the cornea and the lens. The amount of light that enters the eye is controlled by the iris. The light is then translated from biochemical signals into action potentials by the photoreceptors and forwarded through the interneuron to the ganglion cells. The final signal is then transmitted via the optic nerve to the brain.

2.1.2 Visual Function Measures

Besides visual acuity being the most commonly used measure of visual function, there exist other techniques such as contrast sensitivity, visual field test or microperimetry. When performing biomarker discovery in retinal images, it is important to understand the field of application, advantages and limitations of currently used function measures. When evaluating the relationship between biomarker candidates and visual function, it is important to choose an appropriate measurement. For instance, visual acuity remains relatively unaffected in early stages of AMD, while microperimetry is more sensitive in this case (Wu et al., 2014).

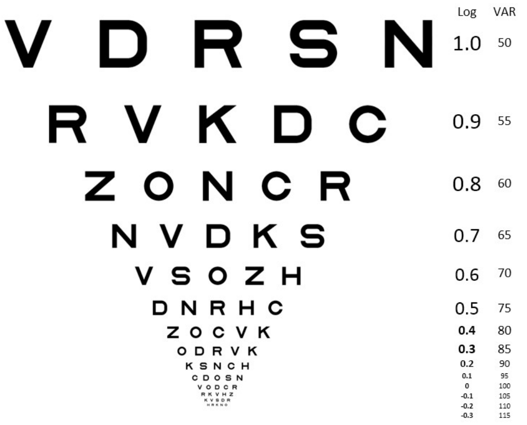


Figure 2.3: The logMAR chart, used to measure best corrected visual acuity (BCVA). Each line contains five letters. The distance between lines as well as the size of letters decreases logarithmically. Source: Commons (2015).

LogMAR	VAR	Snellen (Metric)	Snellen (Imperial)	Dezimal	Number of letters
-0.30	115	6/3	20/10	2.00	100
-0.20	110	6/3.8	20/12.5	1.60	95
-0.10	105	6/4.8	20/16	1.25	90
0.00	100	6/6	20/20	1.00	85
0.10	95	6/7.5	20/25	0.80	80
0.20	90	6/9.5	20/32	0.63	75
0.30	85	6/12	20/40	0.50	70
0.40	80	6/15	20/50	0.40	65
0.50	75	6/19	20/63	0.32	60
0.60	70	6/24	20/80	0.25	55
0.70	65	6/30	20/100	0.20	50
0.80	60	6/38	20/125	0.16	45
0.90	55	6/48	20/160	0.125	40
1.00	50	6/60	20/200	0.10	35
1.30	35	6/120	20/400	0.05	20
1.60	20	6/240	20/800	0.025	5
2.00	0	6/600	20/2000	0.01	–

Table 2.1: Conversion table between different visual acuity scoring systems.

BCVA *Best corrected visual acuity (BCVA)* is the most commonly used measure of visual function (Collin, 2008). In principle, patients are asked to read ”normally” illuminated charts

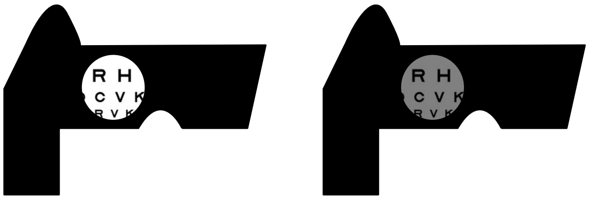


Figure 2.4: Schematic illustration of applying a neutral density filter.

from a specified distance. The term ”best corrected” refers to the fact that refraction of patients is corrected, so hyperopia (farsightedness) or myopia (shortsightedness) should not influence the test result. There are various chart types, for instance a Bailey-Lovie (Bailey and Lovie, 1976) or ETDRS (Ferris III et al., 1982) chart, which are so called logMAR charts (Figure 2.3). The testing distance has to be defined as well: typical distances are 4 or 6 meters (Elliott, 2016). Furthermore, the type of refractive correction must be determined, indicating whether the habitual visual acuity (using the patients own lenses or glasses) or a more specific correction has been used. Another detail that has to be specified is the amount of illumination used for the chart (Ferris and Sperduto, 1982). A so called termination rule is also necessary to achieve a reliable measurement. A typical rule is to stop the patient if four or more errors occur within a line of five letters (Carkeet, 2001). Finally, there exist various scoring systems such as logMAR, visual acuity rating (VAR), number of letters, Snellen or decimal visual acuity. According to Elliott (2016) logMAR should be used, which is based on the total number of letters the patient is able to read. In this logarithmic scale normal vision is represented as 0, while an increased value indicates reduced vision. In particular, each line that can be read corresponds to a change of 0.1 in the logMAR score. VAR and number of letters have been proposed to address the counterintuitive property of negative values in the logMAR score, but conversion and interpretation has to be conducted carefully (Elliott, 2016). The conversion table between different scoring systems is provided in Table 2.1. As one can see, there are many factors that can influence the test result of a visual acuity measurement. BCVA can be unreliable, especially if multiple clinicians conduct measurements without considering a specific standardization system (Elliott, 2016). Hence, it is important to specify and report all details of BCVA measurements (Williams et al., 2008). But even under perfect conditions, patients condition and motivation as well as inter-observer variability can lead to deviating results. It has been reported that BCVA can be ascertained at a confidence level of 95% within ±8 letters (±0.15 log units) under normal clinical conditions (Siderov and Tiu, 1999) and within ±5 letters (±0.1 log units) under optimal conditions (Arditi and Cagenello, 1993).

LLVA *Low luminance visual acuity (LLVA)* is very similar to BCVA. The setting as well as the factors of variation are the same as describe above, except that patients read the ”normal” illuminated chart with a neutral density filter placed in front of their eye that lowers the luminance by a certain factor (Frenkel et al., 2016, Sunness et al., 2008). Figure 2.4 illustrates

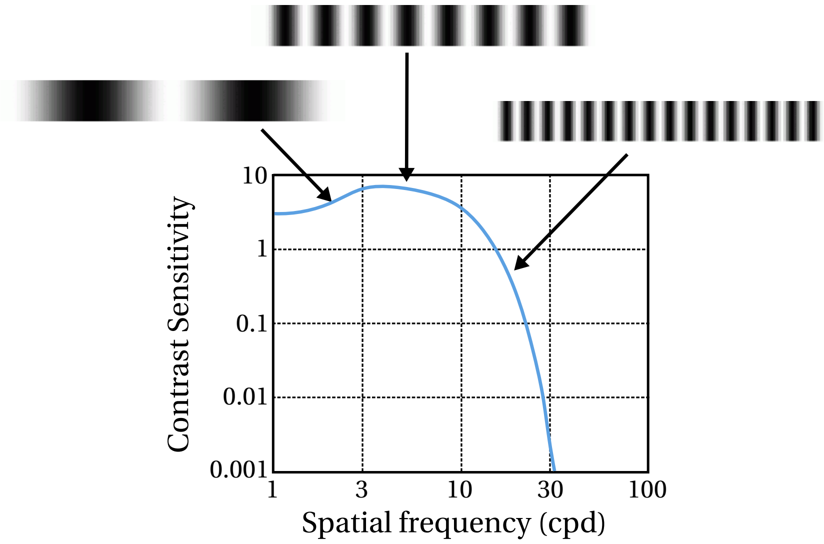


Figure 2.5: Contrast sensitivity function curve according to spatial frequency. Spatial frequencies are measured in cycles per degree (cpd) and influence the amount of contrast needed to distinguish an individual object from its background.

the effect of such a filter.

Contrast sensitivity *Contrast sensitivity* is defined as the ability of distinguishing an object from its background, depending on the contrast between them (Richman et al., 2013). Contrast is the difference between foreground and background in terms of luminance. A gray object against a gray background has low contrast, while a black object in front of a white background has high contrast. The amount of contrast needed to distinguish an object from its background also depends on the object size. This can be expressed in terms of spatial frequency, as illustrated in Figure 2.5. Low spatial frequency is shown as sparsely packed lines, meaning a lower number of lines (circles) with a defined visual angle (degree). Contrast sensitivity can be tested using letter or grating charts. Commonly used tests involving charts are Pelli-Robson (Trobe et al., 1996), Regan (Regan and Neima, 1983), Arden plates (Arden and Jacobson, 1978), Cambridge gratings (Fahy et al., 1989) or Vector Vision (BÜhren et al., 2006) charts. In contrast to visual acuity tests with logMAR charts, these tests evaluate vision at different contrast and spatial frequency levels. A patient may have perfect BCVA and still may suffer from poor vision due to bad contrast sensitivity, affecting the quality of life (Owsley and Sloane, 1987, Richman et al., 2013). For instance, it has been shown that contrast sensitivity is affected by AMD (Faria et al., 2015). While chart based testing faces similar drawbacks as mentioned above for BCVA, there are other challenges for computer based contrast sensitivity examinations. While it allows to show contrast targets in random order or to use staircase strategies, the costs and expenses are higher and test results can be affected by monitor properties such as size or luminance (Richman et al., 2013).

Perimetry The visual field is defined as the perception area of a single eye that is focused on a stationary object, without moving the head or eye (Council et al., 2002). While visual acuity tests measure the high-resolution vision in the center of the visual field, the systematic examination of visual field function is called *perimetry*. It evaluates the ability of recognizing stimuli in the macular, central and peripheral visual field. Perimetry tests can be categorized into *kinetic* and *static automated* perimetry. In kinetic perimetry, the stimulus is moved while size and brightness stay constant. In static automated perimetry, brightness and/or size of the stimulus are varied while remaining at the same position (Council et al., 2002). Perimetry can be particularly useful in the context of diseases that do not affect central but peripheral vision. However, this conventional form of perimetry is not suited to test retinal sensitivity in the macula in a stable and accurate way, especially in patients with unstable fixation (Acton and Greenstein, 2013).

Microperimetry Microperimetry allows to overcome some limitations of automated perimetry, enabling a detailed evaluation of macular function (Acton and Greenstein, 2013, Markowitz and Reyes, 2013). It is the most comprehensive test of the macula regarding visual function, examining retinal sensitivity at specific locations in a non-invasive way. Stimuli are projected directly onto the retina, following a pre-defined grid (Markowitz and Reyes, 2013). This examination is conducted with modern devices such as Macular Integrity Assessment (MAIA^{TM}; CenterVue, Padova, Italy) or Nidek Microperimetry-3 (MP-3; Nidek Technologies Srl, Padova, Italy) (Balasubramanian et al., 2018, Hirooka et al., 2016). These devices use eye-tracking to ensure a certain amount of test-retest stability, meaning that a stimulus that is repeatedly sent with the same intensity to the same position should always reach the same correct location in the retina, leading to the same test outcome. Despite the opportunities of microperimetry, it also faces some issues. The examination takes several minutes and can be exhausting under certain conditions, e.g. large grids or specific disease patterns. The quality and therefore the significance of the test results do not only depend on the technical specifications of the device (Balasubramanian et al., 2018, Hirooka et al., 2016), but also on the motivation, learning effect or fatigue of the patient (Wong et al., 2017). Furthermore, depending on the disease and its concrete manifestation in the patient, the reliability of the test results can also vary (Molina-Martín et al., 2016, Wong et al., 2017, Wu et al., 2013, 2015).

2.1.3 Retinal Diseases

While there are various diseases of the eye that manifest themselves in the retina, we will focus the introduction on age-related macular degeneration (AMD) and diabetic retinopathy (DR), which are among the leading causes of blindness. Though our experimental evaluations are conducted on AMD patients, the developed methods can easily be adapted to work on other diseases such as diabetic retinopathy. In addition, our methods are developed to work

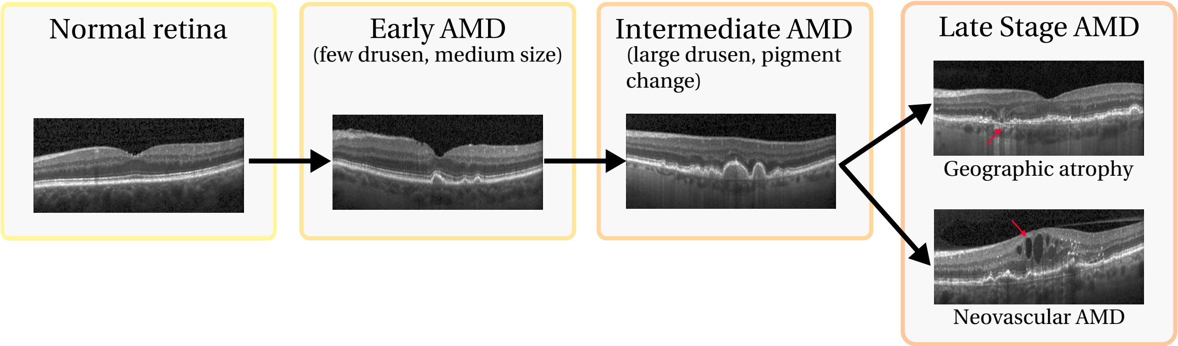


Figure 2.6: Progression paths of age-related macular degeneration (AMD) are depicted. Exemplary optical coherence tomography (OCT) images are shown to illustrate typical visual appearances of different disease stages, from left to right. Early AMD progresses to intermediate AMD as drusen grow and pigment changes occur. In the final stage of the disease, patients can develop both geographic atrophy (GA) or neovascular AMD. The red arrows indicate an area of RPE loss in GA and intraretinal fluid in neovascular AMD, respectively.

on OCT images, which play a crucial role in the diagnosis and management of patients both in AMD and DR.

Age-related Macular Degeneration AMD causes progressive damage to the macula, the part of the retina which is responsible for central high-resolution vision. Therefore, a typical symptom of AMD is loss of central vision. AMD is a common disease, as it is expected that 196 million people will be affected by age-related macular degeneration in 2020, increasing to 288 million worldwide in 2040 (Wong et al., 2014). Figure 2.6 shows the general progression path of AMD. With increasing age, the structure of BM changes and debris from surrounding tissues accumulate above and within the BM. According to Nowak (2006), the impairment of RPE cell function is an early event as well, causing a progressive accumulation of lipofuscin which is composed mostly of lipids and proteins. The aggregation of lipids and other material between the RPE and BM is called drusen and is more prevalent in the macula (Provis et al., 2005). It is important to mention that according to the Beckman Initiative Classification, the appearance of small drupelets only (drusen below 63 μm in size) is considered as normal age-related change, with no clinically relevant increased risk of developing late AMD (Ferris et al., 2013). In contrast, if drusen exceed this size, they should be considered as sign of *early AMD* (Figure 2.6). This small deposits can cause activation of the immune system and local inflammatory response. It is suspected that the activation of the immune system actively contributes to drusogenesis (Nowak, 2006). When drusen get larger and/or changes in the RPE become visible, this is referred to as *intermediate AMD* (Awh et al., 2017). At this point, two subgroups of final disease progressions can be clinically distinguished: *neovascular AMD* and *geographic atrophy (GA)*. While the former is also referred to wet AMD, dry AMD is defined as the absence of neovascularization at any stage.

In neovascular AMD, one hypothesis is that local inflammation causes an increased produc-

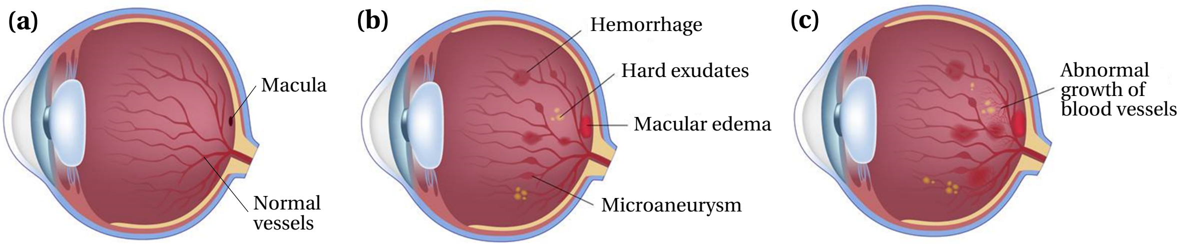


Figure 2.7: Stages of diabetic retinopathy (DR) are shown: (a) no DR, (b) nonproliferative DR and (c) proliferative DR. Modified from source: Orlando (2017).

tion of factors that promote vessel growth, such as vascular endothelial growth factor (VEGF). This means that the balance between pro-angiogenic factors (e.g. VEGF) that stimulate the process of vessel growth and anti-angiogenic factors (e.g. pigment epithelium derived factor (PEDF)) that inhibit angiogenesis, is disturbed (Bhutto et al., 2006). Thus, abnormal vessels start to grow from the choroid towards and through the BM, called choroidal neovascularization. These new vessels are abnormally leaky which causes fluid to enter the retina, with varying degrees of destruction and distortion of retinal structures (Figure 2.6). Besides local inflammation, other factors such as hypoxia may play a role in the development of neovascular AMD as well (Nowak, 2006).

In contrast, the second form of late AMD–geographic atrophy–is characterized by the death of RPE, photoreceptors and/or choriocapillaris. The loss of these structures can occur in arbitrary order as well as in multiple locations at the same time (Boyer et al., 2017, Ferris et al., 2013, Schatz and McDonald, 1989). Moreover, there exists a high interindividual variability, e.g. the progression rates of affected areas vary widely among patients (Ferris et al., 2013, Schatz and McDonald, 1989). The concrete pathogenesis, the differentiation of sub-types regarding risk as well as disease progression are not fully explored and actual topics of research.

In general, the pathogenesis of AMD is complex and multifactorial, involving metabolic, functional, genetic and environmental interactions which are not yet fully understood (Klein et al., 2005, McHarg et al., 2015, Nowak, 2006).

Diabetic Retinopathy As AMD, also diabetic retinopathy (DR) can cause severe vision loss or blindness. It is a common disease with approximately 93 million people affected worldwide (Yau et al., 2012). Symptoms include blurred vision, faded colors, spots or metamorphopsia (straight lines might appear curved). It can emerge in patients suffering from diabetes, where improper metabolization of glucose causes hyperglycemia, defined as an increased level of sugar in the blood. Hyperglycemia can cause pathological changes of retinal vessels. The earliest stage of the disease is called nonproliferative DR and is characterized by the occurrence of microaneurysms, hemorrhages and hard exudates (Abràmoff et al., 2010). Microaneurysms appear when the increased sugar level weakens the vasculature, its rupture

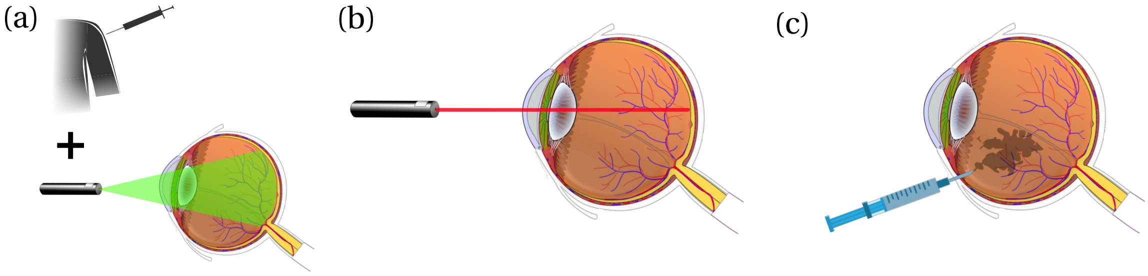


Figure 2.8: Treatment options for wet AMD. (a) In the photodynamic therapy (PDT) a photosensitive substance is administered to the patient, before a low-intensity laser activates it in the abnormal retinal vessels. (b) A high intensity laser surgery is done to locally heat, seal and destroy abnormal vessels. (c) In the anti-VEGF treatment angiogenesis inhibitors are injected into the eye. The current standard therapy for AMD is (c), while (a) and (b) are only used under special circumstances. Figure elements were used from Commons (2018b).

producing leakages of blood, known as hemorrhages (Mookiah et al., 2013). On the other hand, hard exudates are accumulations of proteins and lipids which leave the vessels due to the increased permeability of its walls. However, nonproliferative DR is characterized by the absence of neovascularization. The damage of retinal vessels can cause increased production of pro-angiogenic factors (e.g. VEGF), leading to growth of additional abnormal vessels which show increased permeability. This stage where neovascularization occurs is called proliferative DR. In this context, the abnormal appearance of fluid in the retina is referred to as *diabetic macular edema (DME)* and can occur independent of the DR stage.

2.1.4 Treatment Options and Challenges

Currently, there are three therapies for neovascular AMD available: photodynamic therapy (PDT), laser surgery and injection of angiogenesis inhibitors into the eye (Wykoff et al., 2018). While the former two are only used in specific circumstances, the latter is the currently used standard therapy. In the **photodynamic therapy** a photosensitive substance (verteporfin) is administered intravenously. When this substance has reached the leaking blood vessels in the retina, it is activated by directing a low-intensity laser to the retina (Figure 2.8(a)). This leads to a destruction of the abnormal blood vessels (Al-Zamil and Yassin, 2017, of Age-Related Macular Degeneration with Photodynamic Therapy , TAP). Since the drug is activated by light, it is important not to expose the eye or the skin to sunlight for a few days after the treatment. Moreover, the use of PDT in wet AMD has declined due to inadequate and unpredictable effects that have caused large number of recurrences and retreatments needed (Al-Zamil and Yassin, 2017, Michels et al., 2006).

Conducting a **laser photocoagulation surgery** is another option of treating wet AMD (Wykoff et al., 2018) (Figure 2.8(b)). A high-energy laser is applied with a specific amount of energy to selected locations of the retina to heat, seal and destroy abnormal vessels (Joussen and Bornfeld, 2009). An essential drawback is the blind spots produced by scarring at the area of utilization, leading to a permanent vision loss. As a consequence this therapy is only considered if the area needed to treat is not located in the macula. Moreover, it has a high recurrence rate of almost 50%, meaning that multiple laser surgeries are necessary if new vessels grow (Hubschman et al., 2009, Joussen and Bornfeld, 2009).

The third treatment option is the current state-of-the-art therapy: injections of angiogenesis inhibitors into the eye, also known as **anti-VEGF treatment**. As discussed above in Section 2.1.3, it is known that an increased amount of vessel growth factors (VEGF) leads to angiogenesis of abnormal vessels. A consequence is the leaking of blood and/or fluid into the retina which causes destruction of retinal structures. To prevent the formation of new blood vessels and to reduce vascular leakage the inhibitory substance is directly injected into the vitreous, blocking the activity of VEGF (Figure 2.8(c)) (Wykoff et al., 2018). It allows to slow down the disease progression and can lead to a partial recovery of vision. However, since a single injection costs more than 1000€ and regular treatments are needed at the same time, the consequence is a high burden both for the patient and the healthcare system (Stein et al., 2014). Treatment strategies such as *pro re nata (PRN)* or *treat-and-extend* try to reduce the number of visits needed. After an induction phase of three months with injections at a 4 week interval, PRN treatments are conducted in case of recurring fluid or vision loss (Wykoff et al., 2018). This incorporates the disadvantage of treating a patient after damage has already occurred. *Treat-and-extend* also involves an induction phase, after which the treatment intervals are prolonged at each visit without event (no vision loss and no recurrence of fluid). This forms an incremental approach with increasing time intervals between treatments until signs of CNV reoccur (Wykoff et al., 2018). The limitations of this strategy are two-fold: In the beginning, some patients might receive more injections than actually needed, while at a later stage prolongation of treatment intervals will lead to disease progression and damage of the retina. Due to individual disease progression paths and differing treatment responses, a single treatment strategy is suboptimal (Al-Zamil and Yassin, 2017). In this context, new biomarkers can help to identify patient sub-groups with different risks and treatment requirements at the earliest possible time point. This enables an optimization of treatment for each individual patient, representing a cornerstone of precision medicine.

On the other hand, the investigation of imaging biomarkers is also an important component in the exploration of pathogenesis and disease characteristics, guiding future research and clinical practice (Lambert et al., 2016). Particularly, for the second type of late AMD (GA) there is currently no therapy available which has proven to be effective (Al-Zamil and Yassin, 2017, Holz et al., 2018). Even though it has been shown in the age related eye disease studies (AREDS 1 and 2) that a specific combination of vitamins and minerals for early AMD patients can reduce the risk of disease progression to late AMD by 25% (Group et al., 2000, 2001a), Awh et al. (2015, 2017) discovered that this medication is only effective in certain genotypes, representing an example that emphasizes the need for precision medicine and therefore for biomarker discovery. Moreover, the exploration of biomarkers is needed due to the still existing lack of knowledge regarding the pathogenesis of AMD (Al-Zamil and Yassin, 2017).

For treatment of proliferative DR, panretinal photocoagulation (Chappelow et al., 2012) as well as anti-VEGF treatments are applicable, aiming at stopping proliferation. For diabetic macular edema (DME) anti-VEGF injections, intravitreal corticosteroids or focal/grid photocoagulation can be applied with the goal of stopping leakage and restoring visual acuity (Chappelow et al., 2012, Wenick and Bressler, 2012). Here, anti-VEGF therapies are widely applied, since they have shown efficacy in comparison with laser surgery (Ford et al., 2013). In general, anti-VEGF drugs can be administered in all retinal diseases that lead to macular edema (Channa et al., 2011, Ding and Wong, 2012, Schmidt-Erfurth et al., 2014).

2.2 Optical Coherence Tomography (OCT)

Optical coherence tomography (OCT) is a non-invasive imaging technique in medicine that can produce two- and three-dimensional images. Two exemplary cross-sectional slices of the retina are shown in Figure 2.10. It enables imaging of structures and pathologies at high-resolution level, down to 1-15μm (Drexler and Fujimoto, 2008). OCT has become an important diagnostic modality in ophthalmology, allowing in-vivo diagnosis and assessment of retinal diseases and its progression (Schmidt-Erfurth et al., 2005). For instance, OCT is used to guide different treatment strategies of AMD presented in Section 2.1.4 (Wykoff et al., 2018).

In general, OCT images can be acquired in a non-invasive way, by measuring the reflectivity of tissue types based on low-coherence interferometry (Drexler and Fujimoto, 2008). In contrast to ultrasound (sound waves), MRT (radio frequency) and CT (ionizing radiation), OCT is based on coherent light waves, enabling an acquisition of high-resolution images at μm level. Moreover, it can be safely applied to the retina without causing any damage to the eye. At the same time, the main drawback is the low penetration depth of OCT, restricted to approximately 2mm due to attenuation of light in tissue (Drexler and Fujimoto, 2008).

As illustrated in Figure 2.9, a broadband light beam is emitted from a source. A beamsplitter is used to split the emitted light into two parts: the reference and a measurement beam. While the reference beam is reflected by a mirror at a specific distance, the measurement beam is reflected by the tissue of the retina. The reflected light of both beams is finally combined at the detector. Since the position of the source, beamsplitter and reference mirror are specified, the reference beam travels a known distance at a known time-span until reaching the detector. At the same time, the measurement beam is reflected at multiple depth-levels of the tissue, meaning that reflections at different depths have different travel path lengths back to the detector. However, an interference pattern is only created if the travel paths of the reflected reference beam and the reflected measurement beam differ less than

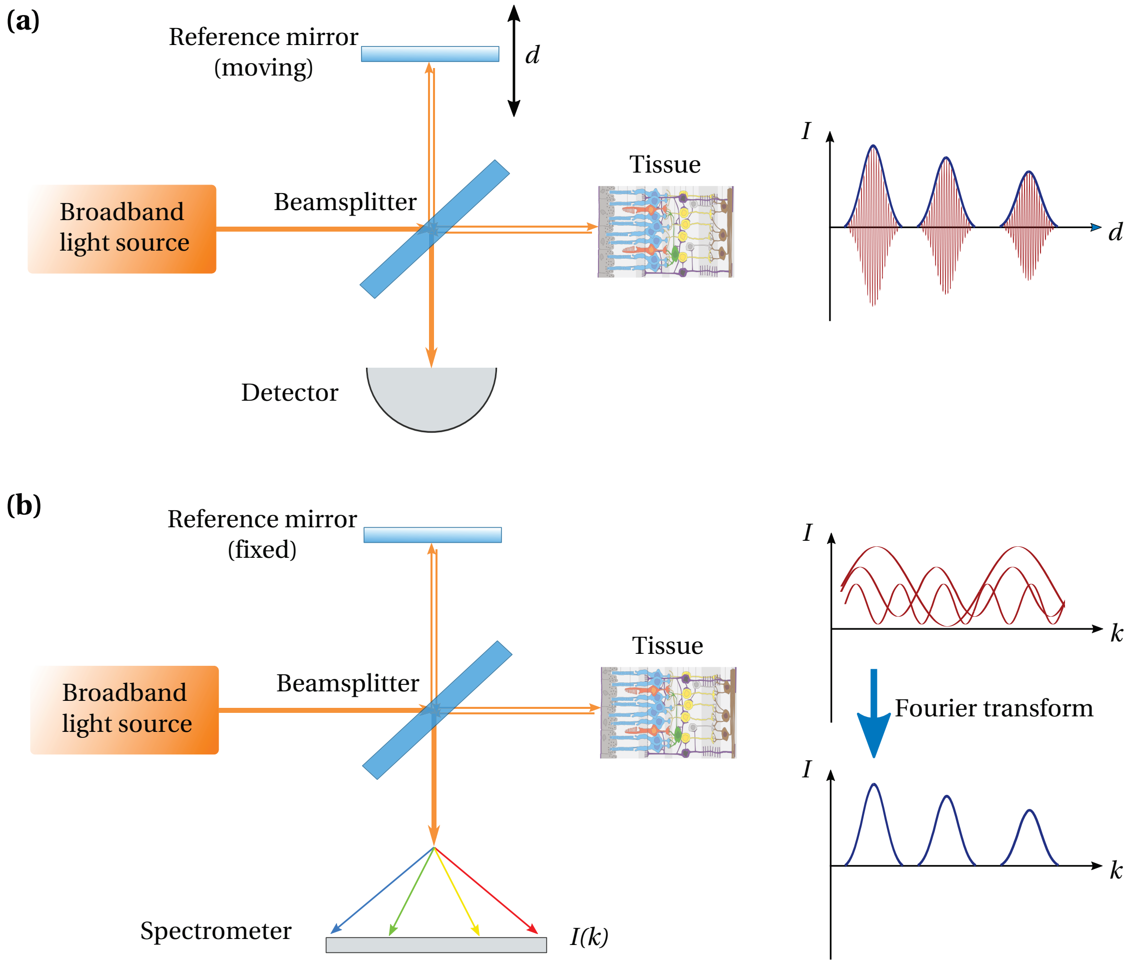


Figure 2.9: Concept of OCT imaging. Schematic illustration of the (a) time-domain and (b) spectral-domain OCT acquisition principles.

a coherence length. Reflections outside of the coherence length do not cause a interference response. The more light is reflected within this coherence length (i.e. at a specific depth of the tissue) the higher is the corresponding interference signal measured at the detector. In other words, the distance of the reference mirror controls the depth at which reflectivity measurements of the tissue are conducted. By translating the reference mirror, interference signals can be measured at different tissue depths. In this way, a one-dimensional reflectivity profile is captured, namely A-scan or depth scan. Multiple A-scans can be combined to form a two-dimensional slice along the lateral axis, called B-scan. A three-dimensional volume (3D-OCT) can be obtained by performing a raster scan, concatenating several B-scans. This type of image acquisition technique is called **time-domain OCT**, since reflections at different depths generate interference signals at different times. However, the scanning speed of time-domain OCT is limited by the speed the reference mirror can be moved (Gabriele et al., 2010) [1]. Besides practical implications for the clinical use case, severe motion artifacts occur

[1] The first commercially available time-domain OCT device had a scanning speed of 400 Hz (400 A-scans/sec), while SD-OCTs reach speeds up to 312.5 kHz (Gabriele et al., 2010, Potsaid et al., 2008)

as a consequence of the long acquisition time.

Several limitations of time-domain OCT are overcome by the novel generation of **spectral domain OCT (SD-OCT)** (Götzinger et al., 2005). According to Drexler and Fujimoto (2008), scanning is 50 to 100 times faster compared to earlier OCT systems. This allows to reduce the effects of motion and to increase the number of acquired A-Scans per slice and/or the number of B-scans per 3D-OCT. Instead of moving the reference mirror to measure intensities at different depths (intensity modulations are measured as a function of time), the position of the reference mirror is fixed and a single A-Scan is acquired in one pass. This is done by using a spectrometer in place of the detector, acquiring the broadband interference (intensities are measured as a function of light frequencies). Applying a Fourier transformation on this spectrometer output yields the reflectivity as a function of depth for a full A-scan (Popescu et al., 2011). A schematic illustration is provided in Figure 2.9

OCT is an essential tool in ophthalmology for diagnosis and management of patients with retinal diseases (Schmidt-Erfurth et al., 2005). Its high-resolution property allows to identify subtle changes in the retina and to detect pathological structures. In clinical practice, OCT devices of different vendors are used, where they differ with respect to imaging and post-processing techniques which leads to varying visual appearance. Figure 2.10 illustrates example B-scans of two different vendors whose OCTs are used in this thesis: Cirrus (Carl Zeiss Meditec, Dublin, CA) and Spectralis (Heidelberg Engineering, Heidelberg, Germany). Besides differing voxel resolutions, the latter involves a process called B-scan averaging during image acquisition. While this leads to an improved signal-to-noise ratio within B-scans, the drawback is an increased scanning time of a single B-scan.

Besides OCT, there exist other retinal imaging modalities such as fundus photography (Panwar et al., 2016) or fluorescein angiography (FA) (Abràmoff et al., 2010). However, in this thesis we develop methods for image biomarker discovery in retinal OCT scans, which is why we limit our description to OCT.

2.3 Clinical Trial Endpoints

Clinical trials are designed to provide substantial evidence for the effectiveness and safety of a specific therapy. The choice of endpoints in clinical trials is important, since they are used to prove this efficacy (Medeiros, 2015). Hence, the endpoints which are used to asses the effectiveness must be clinically meaningful, so the evaluation and the shown effect is meaningful as well. They should enable to show efficacy for patients on average, and at the same time allow to asses the benefit for individual patients (Lassere et al., 2007, Medeiros, 2015).

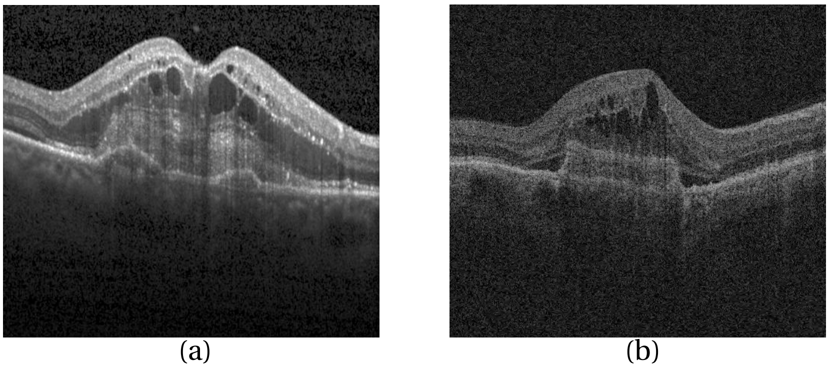


Figure 2.10: Illustration of exemplary spectral domain OCT (SD-OCT) B-scans from two different vendors: (a) *Spectralis* (Heidelberg Engineering, Heidelberg, Germany) and (b) *Cirrus* (Carl Zeiss Meditec, Dublin, CA), with image resolutions of 496×512 and 1024×512, respectively. Both cross-sections show a retina of patients with late neovascular AMD. The differences in image appearance are due to varying image acquisition techniques, e.g. the B-scan averaging technique used in Spectralis devices usually leads to an improved signal-to-noise ratio within slices.

2.3.1 Primary and Surrogate Endpoints

In general, two types of endpoints can be distinguished: **primary** and **surrogate** endpoints. Primary endpoints are patient-centered variables and provide a direct measure on how a patient feels, function or survives. They should represent the clinical outcome of interest themselves and thus being clinically meaningful (Cole et al., 2016). Surrogate endpoints act as a substitute for a clinically meaningful primary endpoint. Anatomical endpoints, laboratory measures or physical signs are examples for surrogates (Cole et al., 2016). An underlying assumption when using them is that surrogate and primary endpoints are related: changes in surrogates are expected to reflect changes in primary endpoints (Prentice, 1989). Before a biomarker can be used as surrogate endpoint in a clinical trial, this expectation needs to be validated. In general, surrogates can be seen as biomarkers that achieved a "surrogate status" in a particular context. In principle, biomarkers are *disease-centered* and reflect the biology and the mechanisms of the disease. By performing validation with respect to a specific primary endpoint, the status of a biomarker moves from a *disease-centered* towards a more *patient-centered* variable. This validation is not a single event, but an incremental process (Lassere et al., 2007). A surrogate that is valid for one primary endpoint, may not be a valid surrogate in another context (Gobburu, 2009). Ideally, the surrogate should be an explicit part of the therapeutic pathway, meaning that the treatment results in the benefit by virtue of its effect on the surrogate (Cole et al., 2016). As described above, primary endpoints have a direct relation to the patient and are defined as the standard to asses therapeuthic interventions. However, reasons for using surrogate endpoints are diverse: they may be faster and easier to obtain, are cheaper or have better reproducibility. Moreover, surrogates can be used as supportive endpoints to better assess the characteristics of therapeutic effects.

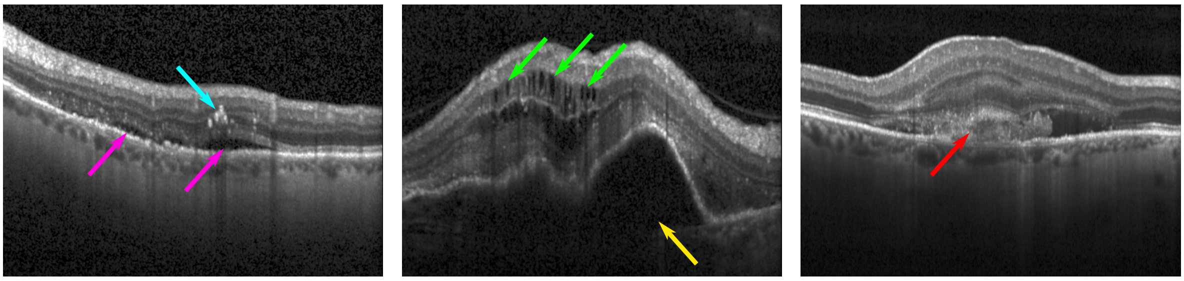


Figure 2.11: Example Spectralis OCT B-scans of patients with neovascular AMD are shown, the arrows illustrating different lesion types (*cyan*: hyperreflective foci (HRF), *purple*: subretinal fluid (SRF), *green*: intraretinal fluid (IRF), *yellow*: pigment epithelial detachments (PED), *red*: subretinal hyperreflective material (SHRM))

2.3.2 Endpoints in Retinal Studies

In retinal studies, variants of **visual acuity** outcome are the most commonly used primary endpoint. Examples for clinical trials that used visual acuity as endpoint are BRIGHTER(Tadayoni et al., 2016), CRYSTAL (Larsen et al., 2016), AREDS (Group et al., 1999), AREDS2 (Chew et al., 2014) or HARBOR (Ho et al., 2014). They involve variants of visual acuity outcome such as difference in BCVA between the first visit and a later time point in the study, or the amount of patients with decreasing VA or increasing VA.

Typical anatomical endpoints are *central retinal thickness (CRT)*, presence or absence of *macular fluid* or the *area of total leakage* (Ho et al., 2014, Larsen et al., 2016, Tadayoni et al., 2016). CRT is assessed by using OCT imaging, calculating the concrete measurements in a semi-automatic way. The grading of macular fluid is also conducted in OCT, where two types of fluid were evaluated in the above mentioned studies: intraretinal fluid (IRF) and subretinal fluid (SRF). These types are distinguished based on the relative spatial location within the retina, as illustrated in Figure 2.11. While SRF is located underneath the neurosensory retina (defined as ranging from the ganglion cell layer (GC) to the outer segment (OS) layer), IRF is located within neurosensory retina. In OCT scans, SRF appears as non-reflective (dark) spaces just above the RPE. In contrast, IRF is characterized by round shapes that are dark due to minimal reflectivity. The area of total leakage is manually assessed using fluorescein angiography (FA) images (Abràmoff et al., 2010).

Other imaging biomarkers of OCT also represent potential new surrogate endpoints for clinical trials. Besides IRF and SRF, some pathological structures that have been described are hyperreflective foci (HRF), PED and SHRM, shown in Figure 2.11. In addition, some of the above mentioned markers have shown limited predictive capability, meaning that there might be other hitherto unnoticed structures or patterns that are still needed to be discovered (Vogl et al., 2017a).

CHAPTER 3

Machine Learning

"Develop a passion for learning.
If you do, you will never cease to grow."

– Anthony J. D'Angelo

In this Chapter we introduce basic machine learning concepts, algorithmic aspects, deep learning fundamentals as well as essential statistical evaluation metrics to assess deep learning methods in the context of biomarker discovery. In Section 3.1 machine learning fundamentals as well as varying paradigms for training are described. Section 3.2 covers a more specific field of machine learning which has recently shown impressive results, namely deep learning. Finally, we describe evaluation metrics in Section 3.3 and provide a summary in Section 3.4.

3.1 Machine Learning Fundamentals

Machine learning is a sub-field of artificial intelligence (AI). In general, AI subsumes all techniques that aim at mimicking human intelligence in an automated way using computers (Pomerol, 1997). This also involves approaches where specific decision processes are explicitly programmed, e.g. with if-then rules. In contrast, machine learning techniques autonomously learn patterns, regularities and characteristics from example data to solve a given task without hand-crafting decision rules explicitly (Hastie et al., 2009). Machine learning is linked with multiple disciplines such as statistics, optimization and computational theory. The general idea of machine learning is that given a specific problem statement, a corresponding dataset and a model with parameters, the dataset is used to optimize the model parameters to solve the problem (Figure 3.1). This results in an optimized model that can be applied in an automated way to new data.

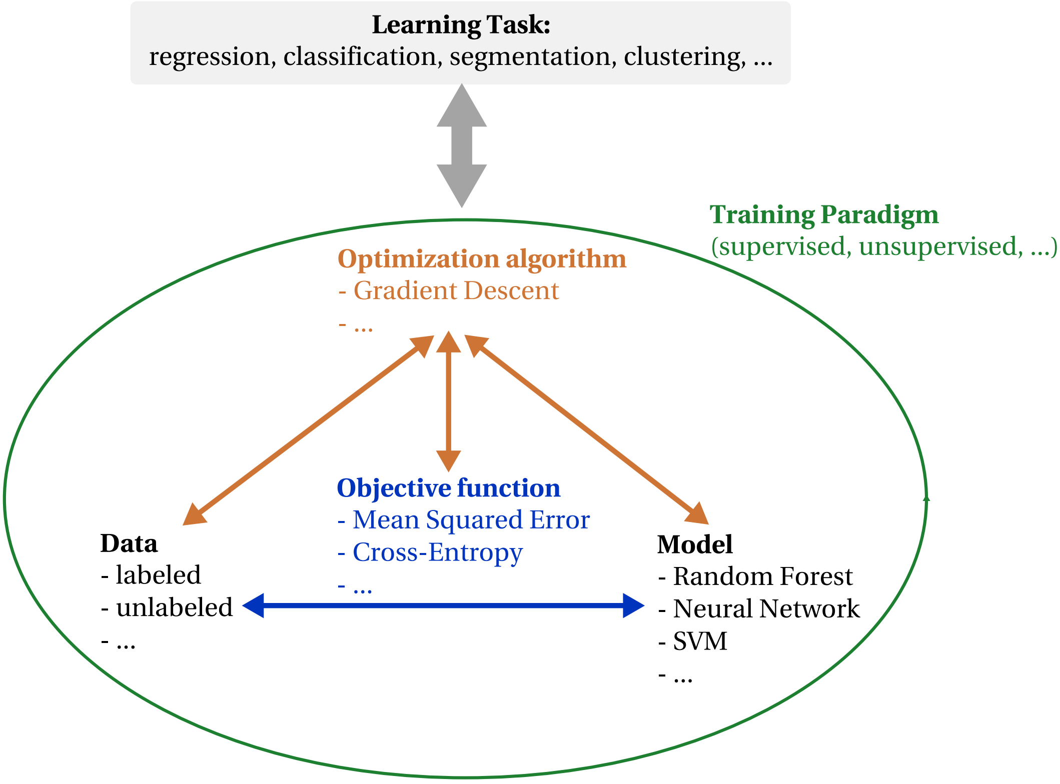


Figure 3.1: Overview of machine learning and its main components. To solve a given task (learning target), an optimization algorithm is applied to optimize the model, using the data and a specific optimization function. The combination of optimization algorithm, data characteristics, optimization function and model choice characterizes the training paradigm.

3.1.1 Training components in machine learning

The process in which the machine learning model is optimized to solve a specific task (*learning target*) is called *training*. The training procedure is also referred to as *learning* and involves (1) training data, (2) a model, (3) an objective function and (4) a specific optimization algorithm. Dependent on the type of model, the characteristics of data, the choice of objective function, the optimization algorithm used and how these components are combined with each other, different types of training paradigms are distinguished from each other (Section 3.1.3). An overview of this setting is given in Figure 3.1.

(1) Data A sufficiently large amount of data is critical for successful training of a model. The amount of needed samples depends on the complexity of the task, the variability of the data, the data quality and the specific model which is trained. The more complex the task, the more variability is present or the lower the quality of the data, the more training data is required. Data is typically split into *training*, *validation* and *test* sets (James et al., 2013). The training of the model is conducted on the training set. The validation- and test sets are not shown to the model during training, since they are used to assess the generalizability of the trained model. In particular, the validation set is used for hyper-parameter tuning and model selection, while the final model evaluation is performed on the hold-out test set. *Cross-validation* is a technique that is especially useful if the amount of data is limited. The data is split into k folds, where one split is used for validation and the remaining ones for training. This is repeated k times, such that each fold is used exactly once for validation (James et al., 2013). In general it is important that the data distribution of the test set matches the distribution of the application domain, so the evaluation on the test set yields a correct estimate of the model performance in the ”real world”.

(2) Model Among others, the given learning task and the available data influence the choice of the specific machine learning model. Examples for different types of models are random forests (RFs) (Breiman, 2001), support vector machine (SVM) (Suykens and Vandewalle, 1999) or neural networks (NNs) (Bishop et al., 1995, LeCun et al., 2015). Each model has specific hyper-parameters (e.g. number of neurons in the NN, number of trees in RFs) which are optimized during training. This corresponds to finding the optimal model complexity. A model that is too complex is likely to *overfit*, having perfect performance on the training set but bad generalization performance on the validation and test set. In contrast, a model complexity that is too low for a given task will face the problem of *underfitting*, showing bad performance on all data sets. Since the final goal is a good generalizability of the model, both cases should be avoided.

(3) Objective function The training of the model is conducted with input samples x and corresponding learning targets y. In general, the model F with parameters θ maps an input x to the output $\hat{y}$:

$$F_\theta(x) = \hat{y}. \tag{3.1}$$

Assuming that for each input x the output targets y are known, the objective function provides a measurement how good the model output is with respect to the known targets. The objective function J(θ) is parameterized by the model parameters θ, since the output of the model depends on θ. A well established objective function is the **mean squared error (MSE) loss**:

$$MSE(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y)^2 \tag{3.2}$$

where n is the dimension of the output. In *classification* problems (Section 3.1.3) a commonly used objective function is the **cross entropy (CE) loss**:

$$CE(\hat{y}, y) = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{3.3}$$

where C is the number of classes, and the target y is provided in an one-hot encoding. This means that y is a vector of length equal to the number of classes, with all elements set to zero except the element that represents the true class which is set to one.

(4) Optimization algorithm The optimization algorithm determines how the model parameters θ are adapted based on the outcome of the objective function $\mathrm{J}(\theta)$ on the training data. The most widely used optimization algorithm for neural networks is *gradient descent* (Ruder, 2016). This technique computes the gradient of the objective function with respect to the model parameters θ, performing a small update of the parameters in order to optimize $\mathrm{J}(\theta)$. A more detailed explanation of NN optimization is provided in Section 3.2.1.

3.1.2 Learning Tasks

There are various categories of learning tasks including classification, segmentation, regression, representation learning or anomaly detection. In the following we provide an non-exhaustive overview of common learning tasks.

Classification In classification, each sample x can be categorized into one out of C classes. In the supervised learning setting (Section 3.1.3), a corresponding target *class label y* exists for each input sample x. **Binary classification** describes the special case in which only two classes have to be discriminated from each other in the dataset. The setting in which the instances are categorized into three or more classes is called **multi-class classification**. Both in the binary and the multi-class setting, a single class label is assigned to each instance. In contrast, if multiple labels can be predicted for each sample this is referred to as **multi-label classification**. Random forests are one example for machine learning classification methods. In Chapter 5, RFs are trained to discriminate patients with different retinal disease stages, based on OCTs.

Semantic segmentation Semantic segmentation describes the task of performing pixel-level detection of objects in an image. While in classification a whole image gets assigned to a single category, in semantic segmentation each pixel gets assigned to a specific class. In other words, the output y is not a single class label, but a label map that is of the same size as the input x. The first group of approaches performs segmentation of the input image in an iterative way, based on classification of patches. Centered at pixel p, a patch $\dot{x}$ is extracted from image x. The machine learning model receives the patch as input and predicts an output class label $\dot{y}$, which is assigned to pixel p. By iteratively doing this for each pixel in the image, the final segmentation map is obtained. The second group of approaches predicts the segmentation map at once, using the whole image as input. A method following this strategy is presented in Chapter 6.

Regression While discrete values are used in classification, regression describes the task of learning the relationship between input data and *continuous* target variables (James et al., 2013). This concept encompasses both the prediction of continuous target values lying within the range of training samples (interpolation) and beyond this range (extrapolation). In Chapter 7, regression is used to evaluate the predictive power of biomarker candidates.

Clustering Clustering describes the process of performing unsupervised (Section 3.1.3) partitioning of data into groups (Jain et al., 1999). Samples within groups are similar, while these similarities are not present across groups. The definition and type of similarity depends on the context and the underlying input data. A widely used clustering technique is *K-means clustering* (Forgy, 1965, Lloyd, 1982). A variant of this algorithm called *spherical K-means clustering* (Hornik et al., 2012) is used in Chapter 5 to identify categories in anomalous areas of OCT scans.

Representation learning This is also known as *feature learning*, targeting to learn a better, more informative representation of the input data, ideally capturing the underlying explanatory factors of the data (Bengio et al., 2013). In contrast to other tasks such as classification or regression, one of the main challenges in representation learning is the difficulty of establishing a clear objective or target to train the machine learning model. However, there are various characteristics of feature representations that are considered to be advantageous, e.g. sparsity, natural clustering or smoothness (Bengio et al., 2013). These assumptions form the basis for many algorithms. For instance, autoencoders (AEs) (Section 3.2.3) belong to the group of methods that learn to map the input data to a lower dimensional *embedding* space, which can be interpreted as data compression procedure. Methods that involve representation learning methods are presented in Chapter 5 and Chapter 7.

Anomaly detection Anomaly detection is sometimes also referred to as *novelty detection* and is defined as the task of identifying test samples that differ in some respect from the data available during training (Pimentel et al., 2014). The normal appearance is learned from the training data, enabling a detection of anomalies that differ from this normal appearance during test time. Hence, the targets of interest (anomalies) are implicitly encoded in the normal training set. *Outlier detection* is a related approach that aims at identifying outliers in the training data (e.g. by fitting models to the region where the training data is most concentrated), and therefore tackles a slightly different task (Chandola et al., 2009). Methods for anomaly detection have been proposed in various domains, such as IT security (Peng et al., 2007), video surveillance (Li et al., 2012, Pokrajac et al., 2007), text mining (Ando, 2007), jet engine monitoring (Hayton et al., 2001) or medicine (Clifton et al., 2011, Schlegl et al., 2017, Sidibé et al., 2017). A widely used technique is the One-Class SVM, aiming at finding a boundary that describes the distribution of normal data and then be used to classify new

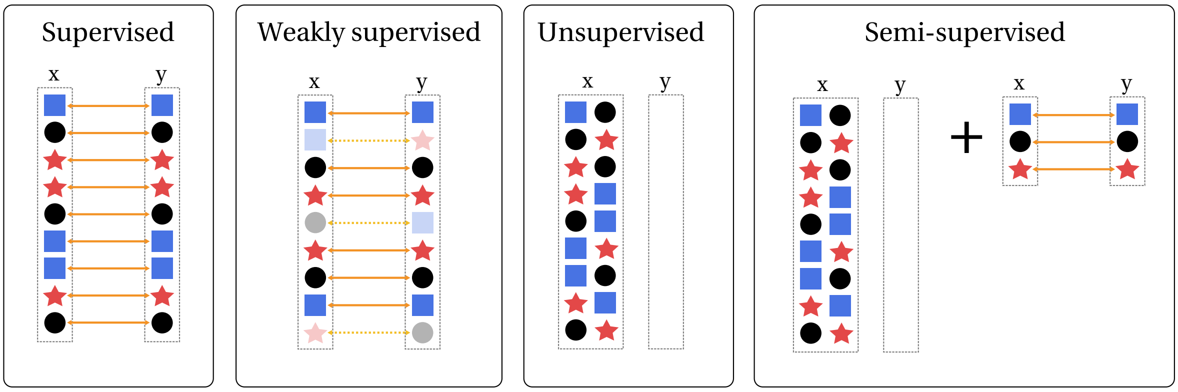


Figure 3.2: Differences between training paradigms are illustrated. In supervised learning, each input sample x has a corresponding target label y. Lower-quality labels are used in weakly supervised learning, meaning that some labels may be incorrect (but we do not know which). In unsupervised learning, models are trained purely data-driven, without labels. Training both on unlabeled and labeled data is called semi-supervised learning.

data as normal or anomalous (Clifton et al., 2011, Hayton et al., 2001, Schölkopf et al., 2001, Tax and Duin, 1999). Another approach for anomaly detection are Gaussian mixture models (GMMs), as used in (Sidibé et al., 2017). The idea is to model the probability density of the normal training data with multiple kernels (Chow, 1970). A drawback of these techniques is the requirement to chose a specific density function, which may not be optimal for the given data (Pimentel et al., 2014). In general, anomaly detection is especially useful in settings where there is insufficient amount of data available to describe the anomalies (Markou and Singh, 2003). Anomaly detection methods are presented in Chapter 5 and Chapter 6.

3.1.3 Training Paradigms

There are different ways how to solve a given learning task with machine learning. In general, the aim of machine learning is to utilize the given training data to automatically find optimal model parameters for a specific task. Based on the information (i.e. data) available during training, we differentiate the following *training paradigms* which are relevant for our thesis: supervised learning, weakly supervised learning, unsupervised learning and semi-supervised learning.

All these training paradigms involve the presence of **input data** x, which represents the *independent variable* or *observed variable*. They differ with respect to the entropy or availability of the **target variable** y (or **label**) during training, which is also referred to as *dependent variable* or *unobserved variable*. An overview of the described training paradigms is provided in Figure 3.2.

Supervised learning In the supervised setting, labeled data is used to train the machine learning model. For each instance x in the dataset, a corresponding *ground-truth* target label y

is available. These labels are generated manually by experts and are provided at the same level of detail as is required by the particular prediction task. For instance, pixel-wise or image-level labels are provided for the semantic segmentation or classification task, respectively. The main drawback of this training paradigm is the fact that manual labels are costly to obtain. Particularly in the field of medical image analysis the acquisition of high-quality labels is challenging, since the annotation process is time consuming and expert knowledge is required.

Weakly supervised learning To overcome the issue that manual annotations are costly to obtain or sometimes not available, weakly supervised learning aims at using higher-level and/or noisier labels as supervisory signal. The idea is to relax the requirement regarding the quality of labels in order to make the labeling process cheaper and affordable. There are various strategies to generate weak labels, such as leveraging higher-level supervision from experts, getting cheaper low-quality labels from non-experts or using already available pre-trained models to produce noisy labels. Higher-level supervision describes the idea of generating labels at a higher level than actually required for the concrete task. In this setting also known as *multi-instance learning*, a set of instances is grouped into a bag, and experts are annotating the bags, not the individual instances (Dietterich et al., 1997, Zhou and Zhang, 2007). Assuming a semantic segmentation problem, experts create image-level (bag) instead of pixel-wise (instance) labels (e.g. information if a tumor is present in the image or not, but no pixel-wise annotations). The image-level (bag) label is then assigned to all pixels (instances), leading to false positive instance labels (but no false negatives).

All weak labeling strategies have in common that they allow to generate labels in a cheap way, at the expense of having incorrect labels for some instances as illustrated in Figure 3.2. A weakly supervised technique is presented in Chapter 6.

Unsupervised learning Unsupervised learning denotes the setting in which no target labels are available. Unsupervised machine learning models are trained without target labels, aiming at capturing the underlying structure of the data (Bengio et al., 2009). Unsupervised learning is closely related to the task of representation learning, transforming the raw input data into a more abstract representation called *embedding space*. Ideally, semantically similar samples have a small distance to each other, while dissimilarity is reflected by large distances in the embedding space. An example for an unsupervised learning algorithm is principal component analysis (PCA), which performs a linear transformation of the input data into a new representation, ranking individual features by their explanatory power (i.e. variance) (López et al., 2011). By retaining only the top-ranked features that contain the highest variance, a low-dimensional embedding space of the input data is established. Since it is a simple but powerful technique, it has been applied for a variety of tasks ranging from medical image compression (Taur and Tao, 1996) to feature extraction (Bogunović et al., 2017). Another technique for unsupervised learning are autoencoders (Bengio et al., 2007), which allow a more complex non-linear embedding and are described in Section 3.2.3. In general, unsupervised approaches can be trained on very large amounts of data, omitting the need for labeling the dataset. Low-dimensional embeddings which have been obtained by unsupervised learning, can be subsequently used as features for clustering or supervised training (*semi-supervised learning*) (Coates, 2012). Unsupervised methods are presented in Chapter 5 and Chapter 7.

Semi-supervised learning Although unsupervised algorithms can be trained without labels on large-scale data, the resulting feature representation may not be optimal for a specific task. In semi-supervised learning, both large amounts of unlabeled data and additional few labeled data samples are used to train the machine learning model (Chapelle et al., 2006). A commonly used strategy involves sequential training on both datasets. First, unsupervised learning is performed on the unlabeled data to learn a mapping to a representative embedding space. Subsequently, the trained model is utilized to map the small labeled dataset to the learned embedding. This feature representation is then used together with the labels to perform supervised training of a second model (Schmidhuber, 2015). Alternatively the parameters of the first model are optimized in conjunction with the parameters of the second model, during the second stage (Schlegl et al., 2014).

3.2 Deep Learning

Since the development of deep learning approaches is the focus of this thesis, a condensed overview about deep learning is given in this section. In particular, main building blocks and principles, as well as supervised, unsupervised and Bayesian deep learning approaches are discussed concisely.

Conventional machine learning techniques are not able to learn features directly from data in an end-to-end architecture. Instead, domain expert knowledge is required to perform careful engineering of feature extractors (*hand-crafted features*). These transform the raw data into a representation from which a learning algorithm can be trained to detect or classify patterns (LeCun et al., 2015, Litjens et al., 2017). An alternative to careful feature engineering is to generate a high-number of feature candidates and perform feature selection in a subsequent step, e.g. with bagging or boosting (Bryll et al., 2003, Langley, 1994, Opitz, 1999)

In contrast, deep learning algorithms automatically learn feature representations optimized for a specific task from data. The term *deep* refers to the fact that the models are artificial neural networks (ANNs) composed of many layers, each layer transforming the input to a slightly more abstract representation. By stacking many layers on top of each other complex functions can be learned, mapping the raw input (e.g. pixel intensity values) to the final output (e.g. class probabilities) (Bengio et al., 2009). For instance, when trained with images, the first layer of a neural network (*input layer*) may learn to detect edges, while

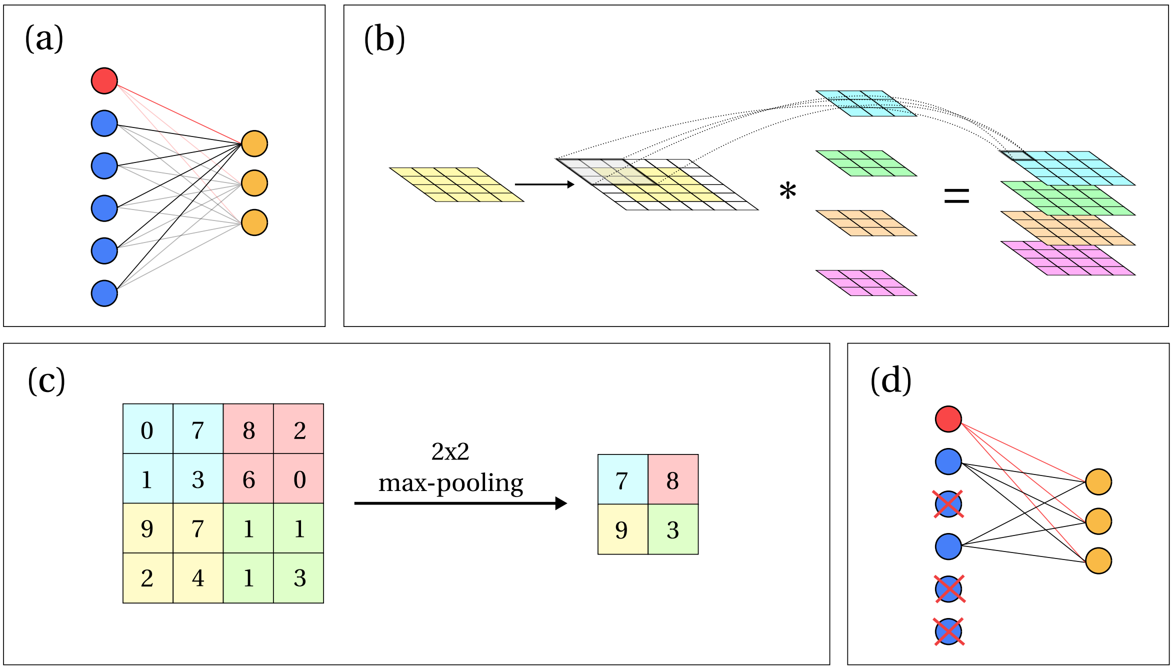


Figure 3.3: Deep learning building blocks. (a) A fully connected layer with an input vector with 5 neurons (blue), the bias term (red) and 3 output neurons (orange). (b) Illustration of a two-dimensional convolutional layer. First, zero padding (black arrow) is applied to a 4x4 input with 1 channel (yellow). The padded input is then convolved with a set of 4 filters $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4\}$ (3×3 filter size, stride of 1), resulting in 4 output channels. (c) Max-pooling with non-overlapping 2×2 blocks. (d) In dropout activations are randomly set to zero with a certain *dropout probability* during training. Here we show the application of dropout to a fully connected layer.

intermediate layers (*hidden layers*) will learn to detect local shapes, object parts and finally entire objects before reaching the final layer (*output layer*). Deep neural networks (DNNs) are inspired by the architectural depth of the brain, specifically by the visual system (Bengio et al., 2009, Serre et al., 2007). Hence, the most basic processing units in DNNs are called *neurons*, sometimes also referred to as *network nodes* or *units*.

3.2.1 Basic elements of deep learning architectures

In general, deep neural networks are composed of various building blocks that form the final architecture of the network. The most fundamental layers are *fully connected* and *convolutional layers*. *Activation functions* are applied to the output of other layers, allowing the network to learn complex non-linear functions. *Normalization* and *regularization* methods are used to improve and stabilize the training of networks. Finally, different *optimization* techniques can be applied to learn the network parameters.

Fully connected layer Fully connected layers can be used in conjunction with non-linear activation functions to learn a non-linear transformation from inputs to outputs, without applying assumptions about the structure of the input features. In a fully connected layer, each neuron is connected to all neurons of the previous layer. The neurons from the previous layer form an input vector x, which is multiplied with a weight matrix $\mathbf{W}$, followed by an addition of a bias term b:

$$z = \mathbf{W}x + b, \tag{3.4}$$

z denoting the output vector of the actual layer, also known as *activations* (Litjens et al., 2017). This constitutes an affine transformation, where W and b are the parameters which are adapted during training. The *number of neurons* in a layer is a hyper-parameter that has to be defined and determines the output size of the layer. Figure 3.3(a) illustrates a fully connected layer with 5 input and 3 output neurons.

Convolutional layer In convolutional layers, each neuron is only connected to a few nearby neurons of the previous layer, exploiting the spatial structure of the input. While fully connected layers can deal with arbitrary dimensions by transforming the input into a one-dimensional representation (vectorization), they have an important drawback. By using vectorized input data, fully connected layers ignore the intrinsic structure of data, such as topological or temporal information. Hence, they represent over-parameterized models when being directly applied to images. In contrast, convolutional layers are parameterized by a set of filters $\mathcal{W} = \{\mathbf{W_1}, ..., \mathbf{W_k}\}$ which are convolved ($*$) with the input X to obtain the activations Z:

$$Z = \mathcal{W} * X + b. \tag{3.5}$$

In other words, each convolution filter $\mathbf{W_j}$ in $\mathcal{W}$ is slid step-wise over the input X, calculating the activations by computing the dot-product between the filter coefficients and the input values it overlaps with. Since the same convolution filter is applied at multiple locations (*weight-sharing*), the parameters of the convolutional layer are reused, reducing the number of parameters compared to a fully connected layer (Litjens et al., 2017). Moreover, the utilization of convolutional filters is motivated by the assumption that filters which are useful in one part of the image should be useful in other locations as well (e.g. edge detectors). In medical image analysis, 2D convolutions are widely applied to 2D images (e.g. OCT B-Scans) (Lee et al., 2017, Schlegl et al., 2018), whereas 3D convolutional layers are applied to 3D volumes (Çiçek et al., 2016, Milletari et al., 2016).

The **filter size** of $\mathcal{W}$ determines the receptive field of a convolutional filter. To increase the receptive field size, either the filter size can be increased or multiple filters can be iteratively applied, e.g. by convolving the image twice with 3×3 filters a 5×5 receptive field is obtained. The **number of filters** k controls the number of output channels (also termed activation maps or feature maps) of the convolutional layer (Le and Borji, 2017, Schmidhuber, 2015). An increasing number of filters increases the capacity of the network, but also the risk of overfitting. Another hyper-parameter is the **stride**, defining the step size with which the convolutional filter is slid over the image and therefore the spatial size of the output feature maps. The spatial size is also determined by the amount of **zero-padding**, a technique that adds zero values at the border of the input (image or feature maps of previous layer) before performing the convolution operation. This is used to preserve the spatial size of the input for the output, since convolution filters can only be applied at image positions where it fully fits into the image. Without zero-padding, the spatial dimensions would be reduced by $f_s - 1$, where f_s is the filter size.

Figure 3.3(b) shows a 2D gray value image with spatial dimensions of $1 \times d \times d$, a typical filter size of 3×3, zero-padding of 1 and the number of filters being $k = 4$. Since the input image has only one channel, the dimensionality of the weight matrix $\mathcal{W}$ is $1 \times 4 \times 3 \times 3$, generating an output with a dimensionality of $4 \times d \times d$.

Activation functions Since fully connected and convolutional layers only allow a linear transformation of the input, activation functions are needed to introduce non-linearity into the networks, enabling DNNs to learn complex non-linear transformations from inputs to outputs. Typically, each fully connected and convolutional layer is followed by an activation function, allowing the network to learn highly complex functions. Traditional activation functions in neural networks are the **sigmoid** σ or **hyperbolic tangent** tanh function, mapping the input to the closed intervals [0,1] and [-1,1], respectively. They are defined as:

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.6}$$

$$f(x) = \tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}. \tag{3.7}$$

In deep neural networks both activation functions are related to the problem of *vanishing gradients* (Bengio et al., 1994). This describes the phenomenon that gradients which are computed to update the parameters of the network *vanish* as the depth of the network increases, leading to slow optimization convergence and poor local optima. The problem of vanishing gradients is addressed by **rectified linear units (ReLUs)** (Hahnloser et al., 2000), the currently most widely used activation function in deep learning (LeCun et al., 2015). It is defined as follows:

$$f(x) = max(0, x), \tag{3.8}$$

mapping the input to its positive part. The gradient used to update the parameters of the network is either 0 or 1 for ReLUs, meaning that the gradient does not saturate/vanish when it is ”transferred” across a network. While ReLUs cause a mean activation greater than zero, **exponential linear units (ELUs)** (Clevert et al., 2015) map the input also to negative values, pushing the mean activation closer to zero. According to Clevert et al. (2015) this leads to faster learning and convergence, having an effect similar to *batch normalization* but with lower computational complexity. It is defined as

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(exp(x) - 1) & \text{if } x \leq 0 \end{cases}, \tag{3.9}$$

where α controls the value to which an ELU saturates for negative values and is by default set to 1.0.

Pooling The pooling layer is also referred to as *downsampling layer*, since it reduces the spatial size of feature maps. Besides the reduction of the spatial resolution, its role is also to introduce translational invariance into DNNs (which is required e.g. in classification tasks). Pooling layers are static, i.e. they do not have learnable parameters. It aggregates sub-regions of input feature maps (i.e. blocks) into single output values by applying a specific pooling operation such as max or average pooling to the input, processing each feature map separately (Boureau et al., 2010a,b). Besides the aggregation type, another hyper-parameter is the block size: a commonly used setting is non-overlapping 2×2 max-pooling blocks that reduce the spatial size of feature maps by a factor of 2 (Figure 3.3(c)). Pooling layers do not only reduce the spatial size and therefore the computational costs, but also enlarge the receptive field of the network and introduce invariance to small translations of the input (Saxe et al., 2011). An alternative to pooling layers are *strided convolutions*, as proposed in (Springenberg et al., 2014). This can help to overcome issues of pooling layers when inverting neural networks (Bruna et al., 2013), or to stabilize training of specific methods such as generative adversarial networks (GANs) (Radford et al., 2015).

Batch normalization Batch normalization stabilizes the training of DNNs by normalizing the layer inputs for each *mini-batch*. It has been shown that normalization of neural network inputs helps to improve and speed up training, for instance by applying *whitening*[1] to the input data (LeCun et al., 2012, Wiesler and Ney, 2011). Besides the input, also intermediate representations of neural networks can be normalized, for instance with batch normalization (Ioffe and Szegedy, 2015). Without this normalization, the distribution of layer inputs is constantly changing during training of DNNs, due to adaptions of parameters of the previous layer. This effect known as *internal covariate shift* destabilizes and slows down the training. Batch normalization enables the use of higher learning rates and therefore accelerates the training process. It tackles the covariate shift effect by normalizing the activation outputs of layers via *mini-batch* statistics. A mini-batch describes a set of input samples $\mathcal{B} = \{x_1, ..., x_m\}$ which are jointly used to adapt the model parameters in a single update step. The following

[1] Whitening is a linear transformation that leads to uncorrelated features in the new representation, with zero mean and unit variance.

calculations are performed to normalize inputs from the previous layer:

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2, \qquad \hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}, \qquad z_i = \gamma\hat{x}_i + \beta, \quad (3.10)$$

where $\mu_{\mathcal{B}}$ denotes the mini-batch mean, $\sigma_{\mathcal{B}}^2$ the mini-batch variance, $\hat{x}_i$ the normalized input, γ and β the learnable parameters of the batch normalization layer and z_i the scaled and shifted normalized activations. During test time, the learned parameters γ, β are applied and population-wide statistics are used (i.e. mean and variance are fixed).

Beside its normalizing property, batch normalization also has a regularization effect onto the model during training. This can be explained by the fact that activations for a training example are not deterministic, since they depend on other examples present in the mini-batch (Ioffe and Szegedy, 2015). Further details of the batch-normalization approach can be found in Ioffe and Szegedy (2015).

Dropout Dropout is a widely used regularization technique for DNNs to reduce overfitting. With increasing number of model parameters θ, the capacity and power of the model increases, but at the same time the risk of overfitting increases as well (Srivastava et al., 2014). Overfitting describes the phenomenon that the model does not only learn the underlying structure of data, but also the noise which is present in the training set. This leads to good performance on the training set and to low generalizability (i.e. high generalization error on validation/test sets) at the same time. Regularization techniques help to reduce the generalization error.

A widely used regularization approach is dropout, where activations are randomly set to zero while performing a feed-forward pass during training (Hinton et al., 2012b, Srivastava et al., 2014). A visualization of dropout is presented in Figure 3.3(d). The idea is that co-adaptions of features are minimized, making the individual layers and therefore the overall model robust with respect to random perturbations of the input. Dropout can be applied to convolutional as well as fully connected layers. Every single activation value in the input x is multiplied with a Bernoulli random variable r which takes the value 0 with probability p and 1 with probability $1-p$. The only hyper-parameter of the dropout layers is p, termed as *dropout probability*. Dropout is not applied at testing time, but taken into account by multiplying the layers weights with $1 - p$, ensuring that for each neuron the expected matches the actual output distribution at test time. Dropout can be seen as sampling randomly thinned networks from the basic network during training, while during testing an averaged prediction from these individual "experts" is obtained. In contrast to *bagging* where multiple models are trained on different subsets of data, dropout operates in the feature space, each model is trained for only one step and all of the models share parameters.

There exist also other techniques to avoid overfitting. Examples are *DropConnect* (Wan et al., 2013) or adding regularization terms in the objective function (e.g. *weight-decay*) (Krogh and Hertz, 1992).

Optimization The optimization of DNNs aims at learning optimal model parameters for a given task with a specific loss function. The training (or optimization) of deep neural networks can be divided into three steps which are constantly repeated throughout the optimization: (1) forward pass, (2) backpropagation and (3) weight update. Additionally, **intialization of model parameters** is done once before the training starts. Commonly used advanced initialization procedures such as *Xavier* (Glorot and Bengio, 2010) or *He* initialization (He et al., 2015) perform a specific random initialization of the network weights to speed up training.

The training procedure starts with a **forward pass** (or *forward propagation*) of an input sample through the network. Activations are computed layer by layer, starting with the first layer close to the input through until the output layer. The prediction of the network is then compared with the ground-truth target label, particularly by computing a loss (or *prediciton error*) using the task specific objective function $\mathrm{J}(\theta)$ (Section 3.1.1). The **backpropagation** algorithm (Rumelhart et al., 1986) is used to compute the contribution of each unit in every layer to the prediction error. This is done by calculating the partial derivatives of the network parameters with respect to the computed error. Finally, the partial derivatives (or *gradients*) are used to update the model parameters, called **weight update**. The step size of the update in the parameter space (Duda et al., 2001) is determined by a hyper-parameter called *learning rate*.

Various optimization algorithms are based on this principle. *Stochastic gradient descent (SGD)* updates the model parameters after each forward pass of a single input sample. In contrast, *batch gradient descent (BGD)* computes the average gradient over all samples in the training set, which leads to more stable estimations of the gradients, but at the same time slows down the training (number of updates per computation decreases). *Mini-batch gradient descent (MBGD)* constitutes a trade-off between SGD and BGD, since a parameter update is based on the average gradients over a subset of the training data. These three approaches differ in terms of the number of samples which are used to calculate a single update step (Bengio, 2012, Ruder, 2016). Typically, the training set is passed multiple times during training, where one pass is termed *epoch*. Other examples for advanced gradient descent based optimization algorithms that are widely used to train deep neural networks include *Momentum* (Qian, 1999), *Adagrad* (Duchi et al., 2011), *RMSprop* (Hinton et al., 2012a) or *Adam* (Kingma and Ba, 2014). An overview of gradient descent methods can be found in (Ruder, 2016).

3.2.2 Supervised deep learning

The above introduced deep learning building blocks are combined in various ways to solve different learning tasks. A widely used family of architectures can be subsumed under the term **convolutional neural network (CNN)**. Convolutional layers are the core building blocks of CNNs, where the basic structure consists of multiple convolutional layers stacked on top of each other, each followed by an activation function and pooling layers. They reduce the spatial dimensionality of the representation layer by layer and capture the context and relevant features for the given task. In classification tasks, this structure is typical followed by fully connected layers and a final classification layer. Since CNNs encode the important features of an input image, this kind of architecture is also referred to as *encoder*.

One of the milestones in the development of deep learning is the work of LeCun et al. (1989), where the backpropagation algorithm was applied to train a CNN. The first successful real-world application of a CNN was presented in LeCun et al. (1998) for hand-written digit recognition. However, it took three days to train this relatively small network *LeNet-5* (approximately 60 thousand parameters). In this context, Cireşan et al. (2010) introduced a fast implementation of CNNs on graphics processing units (GPUs). Since the training of modern deep learning architectures heavily relies on the computational power of GPUs, this work provided an important contribution to the development of deep learning. With the increasing computational power, superhuman pattern recognition performance could be achieved for the first time in Cireşan et al. (2011, 2012). A similar network known as *AlexNet* was proposed in Krizhevsky et al. (2012). It had more than 60 million parameters, was trained on two GPUs and won the large-scale ImageNet (Deng et al., 2009) challenge, where algorithms are trained on more than 1.2 million images to distinguish 1000 different classes. These architectures used sigmoid or hyperbolic tangent activation functions, had kernels with large receptive fields and were quite shallow, with two to five convolutional layers only. More recent architectures such as *VGG* (Simonyan and Zisserman, 2014), *GoogLeNet* (Szegedy et al., 2015), *Inception-v3* (Szegedy et al., 2016), *ResNet* (He et al., 2015) or *DenseNet* (Huang et al., 2017) have a deeper structure with up to several hundreds of layers, use a smaller filter size and different activation functions such as ReLUs.

Besides classification, **semantic segmentation** is also a common task in both natural and medical image processing. As discussed in Section 3.1.2, segmentation can be done by iteratively classifying individual pixels based on patches, or by predicting the whole segmentation map at once. One drawback of the patch-based approach is that input patches from neighbor pixels have a large overlap and the same convolutions are computed multiple times. Examples for deep learning approaches following this strategy are Ciresan et al. (2012) or Schlegl et al. (2015). In 2015, Ronneberger et al. (2015) proposed the **U-net**, an encoder-decoder architecture that predicts the whole segmentation map for the input at once and is built upon the *fully convolutional net* (Long et al., 2015). The basic architecture is illustrated

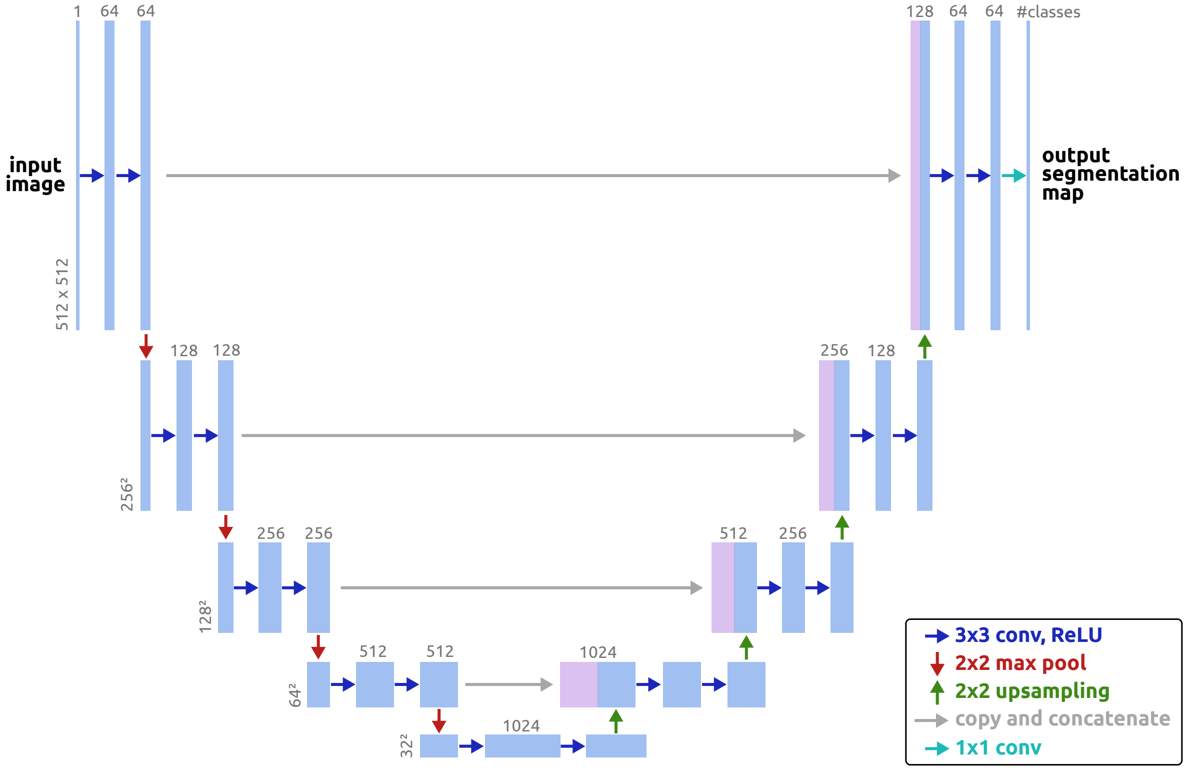


Figure 3.4: Schematic illustration of the U-net architecture. An input image of size 512×512 with one channel is fed to the network. The encoder on the left side contracts the input, while the decoder on the right side recovers the spatial resolution. Blue boxes correspond to the feature map representation, while purple boxes represent copied feature maps. The layers of the network are depicted as arrows. The number of channels in the output layer corresponds to the number of predicted classes.

in Figure 3.4. The encoder follows the typical structure of convolutional neural networks: it consists of convolutional blocks (multiple convolutions with ReLU activation functions), followed by max-pooling layers. The encoder serves as contraction path that reduces the spatial dimensionality, learns high-level abstract features and captures the context. The decoder counterpart performs up-sampling operations of the encoded representations and concatenates them with the encoder feature maps of the next level (same spatial resolution) through shortcut-connections, followed by convolutional blocks. In this way, the contracted information of the encoder is gradually recovered by the decoder (i.e. object details, spatial dimension), allowing the model to obtain a segmentation map that matches the size of the input image. No fully connected layers are used in this architecture. While the original U-net was presented for segmentation of 2D biomedical images, U-net based 3D volume segmentation algorithms have been proposed in Çiçek et al. (2016), Milletari et al. (2016).

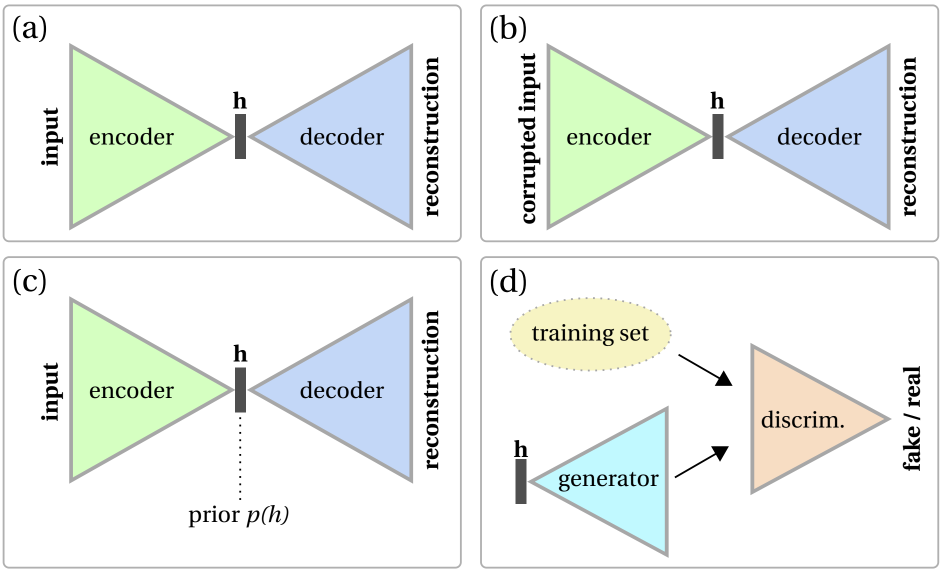


Figure 3.5: Illustration of unsupervised learning approaches. (a) In traditional autoencoders (AEs) the input is mapped to an embedding h using an encoder, aiming at reconstructing the original input through a decoder. (b) The denoising AE reconstructs the original input from a corrupted version in order to get invariant to noise, capturing the underlying structure of the data at the same time. (c) A variational AE puts an additional prior $p(h)$ on the embedding, constituting a generative model. (d) In the training of generative adversarial networks (GANs), a generator learns to generate samples that follow the distribution of the training set, while a competing discriminators tries to distinguish fake from real samples.

3.2.3 Unsupervised deep learning

While impressive results have been achieved with supervised deep learning methods, they rely on *large-scale labeled* datasets. This is a serious limitation especially in medical image analysis, which is one of the reasons why unsupervised deep learning methods have gained attention as well.

Autoencoders (AEs) consist of an encoder and a decoder. Trained without labels, the encoder maps the input x to a lower-dimensional representation h and the decoder reconstructs the original input from this embedding. The difference between the original input x and the reconstruction $\tilde{x}$ is used as error signal to train the AE. A commonly used reconstruction loss is the MSE (Section 3.1.1). The underlying assumption is that the output can only be reconstructed if the learned embedding captures the underlying structure of the data (Baldi, 2012, Erhan et al., 2010). However, if the intermediate representation would have the same size as the input and no non-linear activation functions were added, the model would only ”learn” the identity function (Litjens et al., 2017). To prevent trivial solutions and to enhance the learning process, Poultney et al. (2007) proposed to use *sparsity* constraints on the intermediate representation h. Other work proposed **denoising autoencoders**, where the model is trained to reconstruct the original input from a corrupted version, e.g. by randomly setting values to zero or applying salt-and-pepper-noise (Vincent et al., 2008).

In the simplest form of AEs, the input is encoded and reconstructed with only two consecutive fully connected layers. Deeper fully connected networks are traditionally trained in a greedy layer-wise fashion, where layers are trained and stacked on top of each other consecutively, one after another (Bengio et al., 2007). A different strategy to train deep autoencoders is to optimize all layers jointly at once, which can help to overcome some limitations of the greedy layer-wise scheme (Zhou et al., 2015). Convolutional layers have also been used in the context of autoencoders, where multiple convolutional layers form a **convolutional autoencoder (CAE)** (Masci et al., 2011, Zhao et al., 2016). Autoencoders have also been applied in medical image analysis, e.g. for computer-aided diagnosis (Cheng et al., 2016) or breast density estimation (Kallenberg et al., 2016).

Variational autoencoders (VAEs) as well as **deep belief networks (DBNs)** are also based on the principle of unsupervised representation learning by reconstructing the input. While DBNs can be seen as a composition of restricted Boltzmann machines (RBMs) (Bengio et al., 2007, Hinton et al., 2006), variational autoencoders regularize the representation h to follow a pre-defined distribution via an additional term in the loss function, namely a standard normal distribution with zero mean and unit variance (Kingma and Welling, 2013). Hence, the VAE is a generative model whose regularized, continuous embedding allows to draw random samples from the latent space or generate variations of an input image.

Another family of generative models are **generative adversarial networks (GANs)**, where two networks (generator G, discriminator D) are trained unsupervised in a competing two player game (Goodfellow et al., 2014). The generator aims at generating realistic data samples that follow the training data distribution, while the objective of the discriminator is to differentiate between real samples from the training set and generated samples. At the end of the training, the generator has learned to generate realistic images. The training of GANs can be difficult: the power of both G and D must be balanced during training to allow continuous improvement of both networks, convergence is hard to achieve due to the characteristic of the minimax game and mode collapse is hard to avoid. Several works have been presented to address these issues (Gulrajani et al., 2017, Metz et al., 2016, Radford et al., 2015, Salimans et al., 2016).

In Pathak et al. (2016), the concepts of AEs and GANs are combined with each other in a **context encoder**. A connected region is removed from the input, where the objective of the deep learning model is to reconstruct the original input with a encoder-decoder architecture. In addition to the reconstruction loss, an adversarial loss is provided through the discriminator that distinguishes between original and reconstructed images. The assumption is that the context encoder needs to understand the content of the image and must be able to produce a plausible hypothesis for the missing part in order to succeed at this task. An unsupervised learning algorithm solely based on context information has been presented in Doersch et al. (2015), using the relative position of two patches to each other as training signal.

3.2.4 Bayesian deep learning

Besides the deterministic output of models, providing the uncertainty of predictions is crucial to understand characteristics and limitations of trained models and the underlying task. This is of particular importance in the medical domain, where the model might not be able to generalize to every possible scenario due to the large variety of visual appearances and artifacts. In fact, producing deterministic outputs without reasoning hinders the adoption into clinical routines (Nair et al., 2018). Providing uncertainty estimates for predictions would facilitate subsequent revision by clinicians, referring cases with uncertain diagnosis for further testing (Leibig et al., 2017, Nair et al., 2018).

Bayesian deep learning describes the family of deep learning approaches that can also model uncertainty. It combines the power of deep learning with Bayesian probability theory, enabling the network to infer complex multi-modal posterior distributions. In principle there are two types of uncertainty, namely **aleatoric uncertainty** and **epistemic uncertainty**, which are illustrated in Figure 3.6. The former describes the uncertainty which is related to the information that cannot be explained by the data (Der Kiureghian and Ditlevsen, 2009). For instance, occluded objects or a lack of visual features in images (e.g. because of noise) will cause higher aleatoric uncertainty. The more explanatory variables which are relevant for the given task are present in the data, the lower it is. More precisely, aleatoric uncertainty can be subdivided into *homoscedastic* (task-dependent) and *heteroscedastic* (data-dependent) aleatoric uncertainty. While homoscedastic uncertainty depicts the uncertainty that is inherent in a specific task (staying constant for all input samples and varying between tasks), heteroscedastic uncertainty depends on the input (Kendall and Gal, 2017). In contrast, uncertainty is categorized as epistemic if it can be reduced by observing more data or refining the model (Der Kiureghian and Ditlevsen, 2009, Kendall and Gal, 2017). In other words, it refers to uncertainty that emerges due to informations that in principle could have been captured, but were not. It is also referred to as model uncertainty.

In Bayesian deep learning, epistemic uncertainty is computed by modeling a posterior distribution $p(W|X,Y)$, where W are the weights of the network, X is the training data set and Y the corresponding set of labels. Since finding the true underlying posterior distribution is computationally intractable in practice, it needs to approximated. In Gal and Ghahramani (2015), the authors proposed Monte Carlo (MC) dropout sampling to approximate the posterior distribution with $q(W)$, minimizing the Kullback-Leibler (KL) divergence $KL(q(W)||p(W|X,Y))$. First, the model is trained using dropout. Then, dropout is also applied during test time, which allows to retrieve multiple MC samples by processing the same input multiple times. The mean of the output can be used to obtain a single estimate of the prediction, and the standard deviation of the output reflects an estimate of the epistemic uncertainty of the model (Gal and Ghahramani, 2015, 2016, Kendall et al., 2015).

While Monte Carlo dropout sampling performs variational inference to obtain epistemic

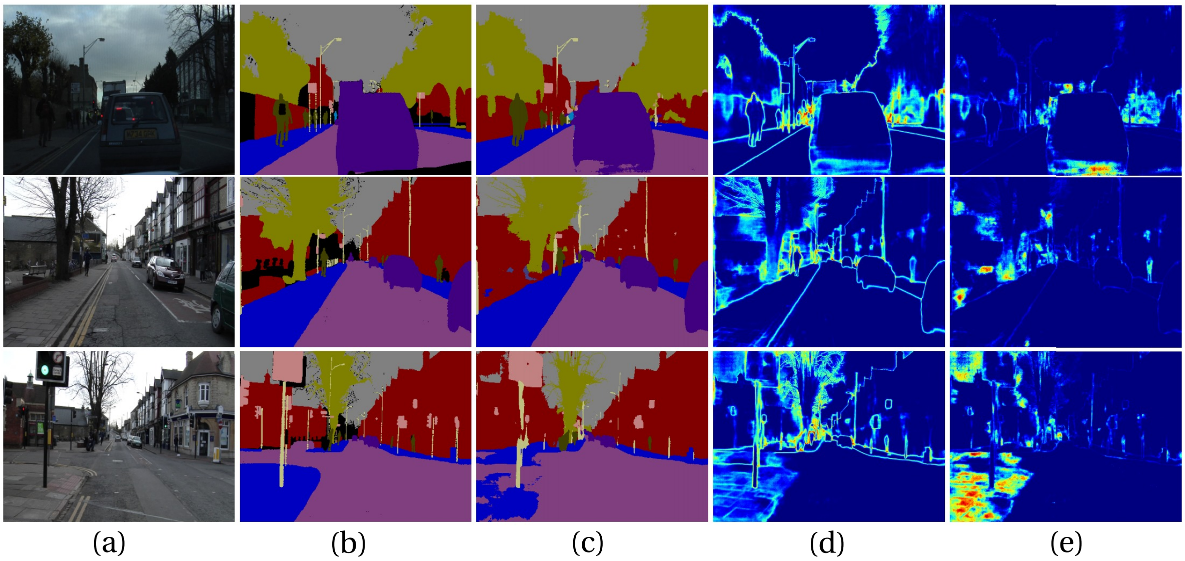


Figure 3.6: Illustration of the difference between *aleatoric* and *epistemic* uncertainty. The (a) input image, (b) ground-truth, (c) semantic segmentation, (d) aleatoric uncertainty and (e) epistemic uncertainty. The aleatoric uncertainty is high at object boundaries and objects which are far away from the camera, capturing the noise inherent in the observations. Epistemic uncertainty captures the model uncertainty which can be reduced by observing ”enough” data. In the bottom row, the model fails to segment the sidewalk due to high epistemic uncertainty. Training the model with more samples of similar appearance could reduce epistemic uncertainty. Figure used with permission of the rights holder: Kendall and Gal (2017).

uncertainty estimates, a maximum a posteriori probability (MAP) inference technique is proposed in Kendall and Gal (2017) to model heteroscedastic aleatoric uncertainty, by changing the loss function of the deep learning model. Instead of only predicting $\hat{y}$, also the observation noise σ is predicted by the model, representing an estimation of aleatoric uncertainty. Considering a loss function $||y - \hat{y}||_2$, the adapted loss according to Kendall and Gal (2017) is defined as:

$$\frac{||y - \hat{y}||_2}{2\sigma^2} + \frac{1}{2}\log\sigma^2. \tag{3.11}$$

The model will learn to attenuate the first term by higher uncertainty σ^2 in cases of bad predictions $\hat{y}$. At the same time, the second term regularizes the magnitude of uncertainty.

3.3 Performance measures

To evaluate the performance of a trained model in a quantitative way, various measures are available. In classification tasks, the following statistical measures are commonly used: **accuracy**, **sensitivity**, **specificity** and **precision** (Sokolova and Lapalme, 2009). All these measures can be described using the four possible outcomes depicted in the confusion matrix of Table 3.1: the number of *true positives (TP)*, *false positives (FP)*, *false negatives (FN)* and *true*

Table 3.1: Confusion matrix of possible test outcomes.

		True value	
		Positive	Negative
Prediction	Positive	*True positives (TP)*	*False positives (FP)*
	Negative	*False negatives (FN)*	*True negatives (TN)*

negatives (TN). Each of these individual statistical measures provides a specific view on the performance. *Combined evaluation measures* such as the **receiver operating characteristic (ROC) curve**, **Sorensen-Dice index** or **precision-recall (PR) curve** allow a joint evaluation of two complementary measures. In regression-tasks, evaluation measures such as the **mean absolute error** or **coefficient of determination** can be used.

Accuracy Accuracy describes the proportion of correct test results:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.12}$$

It is only an appropriate performance measure if the number of samples between classes are balanced, since accuracy will be biased towards the prominent class in an imbalanced setting. This measure is therefore less conclusive in medical image analysis, where the class of interest (e.g. a pathological structure) is often underrepresented (Litjens et al., 2017).

Sensitivity Sensitivity is also known as *recall* or *true positive rate (TPR)*. It is defined as the proportion of true positives which are actually predicted positive. Therefore it describes the ability of the model to detect samples associated with a positive ground-truth label:

$$\text{sensitivity} = \text{recall} = \text{TPR} = \frac{TP}{TP + FN}. \tag{3.13}$$

Specificity The amount of correctly identified negative samples is determined by specificity (or *true negative rate (TNR)*):

$$\text{specificity} = \text{TNR} = \frac{TN}{TN + FP} \tag{3.14}$$

Precision Precision (or *positive predictive value*) is the proportion of predicted positives which are actual positive:

$$\text{precision} = \frac{TP}{TP + FP}. \tag{3.15}$$

Receiver operating characteristic (ROC) curve Sensitivity or specificity should not be interpreted isolated from each other. For instance, if a classifier assigns negative labels to all instances, specificity is 1 while sensitivity is 0. The trade-off between these two measures

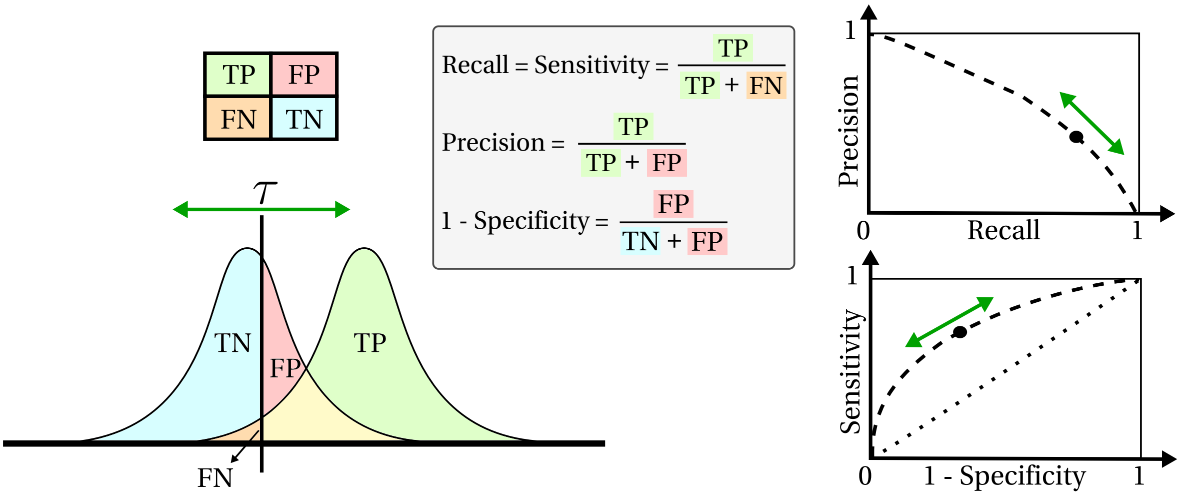


Figure 3.7: Relationship between the model output and the evaluation curves. On the left, the two density functions denote the distribution of positive and negative instances with respect to the output of the model (e.g. probability of assigning the positive class), separated by a threshold τ. The ROC and precision-recall curves are shown on the bottom and top right, respectively.

can be visualized with the ROC curve by plotting the sensitivity against the false positive rate (FPR), defined as:

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{TN + FP}. \tag{3.16}$$

The curve is generated by calculating sensitivity and FPR values at varying threshold (τ) values of the model output, e.g. the probability of assigning the positive class (Figure 3.7). The better the underlying model, the closer the ROC curve is to the top left corner, while the diagonal line corresponds to random performance. Area under ROC curve (AuC) is one way to summarize of the ROC curve in a single value (Sokolova and Lapalme, 2009). It can be interpreted as the average value of sensitivity for all values of specificity, or as aggregated classification performance (Hajian-Tilaki, 2013). In general, the ROC curve is not a good measure in the setting of highly imbalanced data, since the FPR is not sensitive to false positives if the total number of negative instances is huge.

Sørensen-Dice index The Sørensen-Dice index is also known as *Dice score*, *Dice coefficient*, *F_1-score* or *F-measure* and combines precision and recall in one single number (Dice, 1945, Sørensen, 1948). It is defined as

$$\text{Dice} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \tag{3.17}$$

where A and B are two binary sets of ground-truth and predictions, with $|A|$ and $|B|$ the numbers of positive elements in each set. Compared to AuC, the Sørensen-Dice index is better suited for imbalanced class settings, since it is sensitive to the number of false positives and independent from the number of true negative instances (Fawcett, 2004). Due to this characteristic and the fact that segmentations can be interpreted as a set of classifications, it is often used to evaluate segmentation results (Taha and Hanbury, 2015).

Precision-recall (PR) curve Instead of calculating a single value (Sørensen-Dice index), the precision-recall curve provides a graphical illustration of model performance. In contrast to the ROC curve, it shows the trade-off between *precision* and sensitivity (or recall), what makes it the more reliable performance visualization technique in the case of imbalanced classes (Saito and Rehmsmeier, 2015).

Mean absolute error (MAE) In regression tasks, MAE represents a commonly used technique to evaluate the performance of continuous predictions. It has a clear interpretation and is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{3.18}$$

where y_i and $\hat{y}_i$ are the ground-truth and prediction values, respectively.

Coefficient of determination The coefficient of determination is also known as *goodness of fit* or R^2. Ranging from 0 to 1, it describes the proportion of variance in the data that is explained by the regression model (0% to 100%) (Draper and Smith, 2014). It is defined as:

$$R^2 = 1 - \frac{\text{variation of residuals}}{\text{variation of } y} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.19}$$

where $\bar{y}$ is the mean of the ground-truth target values, SS_{tot} the total sum of squares and SS_{res} the residual sum of squares.

3.4 Summary

In this chapter we have described machine learning as well as deep learning fundamentals, forming the methodological background of this thesis. In particular, we covered learning tasks, training principles and basic architectural elements of deep learning models which are used in Chapter 5, Chapter 6 and Chapter 7, forming the context for these presented approaches.

All three manuscripts present *deep learning* approaches, meaning that they are built upon basic elements of deep learning such as *convolutional* and *fully connected layers*, *activation functions*, *pooling*, *batch normalization* and *dropout*. In particular, Chapter 5 proposes an *anomaly detection* method based on *unsupervised learning*, aiming at *semantic segmentation* of anomalies in retinal OCT images. A subsequent *clustering* of anomalies is used to identify marker candidates. In contrast, a *weakly supervised learning* technique is proposed in

Chapter 6, detecting anomalies by using a *U-net* model jointly with *Bayesian deep learning*. Furthermore, the *unsupervised representation learning* technique presented in Chapter 7 is both experimentally evaluated in a qualitative and quantitative way, e.g. in a *regression* task.

CHAPTER 4

Image Biomarker Discovery using Machine Learning

"The best way to predict the future is to create it."

– Abraham Lincoln

Biomarker discovery describes the process of identifying patient characteristics that reflect the biology and the mechanisms of the underlying disease (Group et al., 2001b, Lassere et al., 2007, Medeiros, 2015). As discussed in Section 1.1, it plays a crucial role for precision medicine. Powerful biomarkers can improve the accuracy of diagnoses or allow to detect diseases at an earlier stage. This helps to reduce the burden on the patients and the medical system, since early diagnosed patients can often be treated easier than patients which are diagnosed too late. In the context of retinal diseases, the need for new expressive biomarkers to improve individual patient health care is also highlighted by the fact that the number of patients suffering from AMD is expected to increase, e.g. the number of late AMD cases is estimated to almost double by the year 2040 (Colijn et al., 2017).

Moreover, biomarkers can inform our understanding of the pathogenesis of a specific disease. For instance, specific pathologic changes which are captured by medical imaging could act as risk factors, facilitating the distinction of patients into different risk groups (Cho et al., 2012, Schrag et al., 2000). Biomarkers also pave the way to differentiate individual disease progression paths (Paulovich et al., 2008, Yu and Hung, 2000), as highlighted in Figure 4.1. This means that the exploration of new biomarkers enables an optimization of existing treatments and patient management by both assessing the risk and tracking the disease progression at a finer level (Hughes et al., 2006, Nalejska et al., 2014). Additionally, the gained knowledge about disease pathogenesis can be used to develop new treatments and drugs (Perlis, 2011). Finally, novel biomarkers can act as new surrogate endpoints in clinical

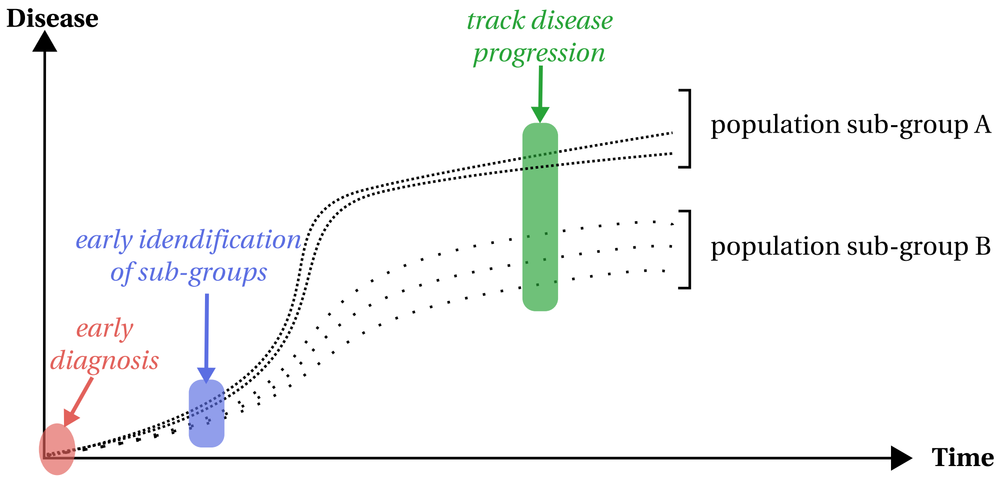


Figure 4.1: Schematic illustration of disease progression paths. Patients with similar progression paths form sub-groups. For instance, these can be characterized by varying medication needs or treatment responses. Expressive biomarkers are necessary in order to allow early and accurate diagnosis (red), early identification of patient subgroups (blue) and tracking of disease progression (green).

trials, allowing to assess the efficacy and characteristics of a therapy in a more comprehensive way (Lassere et al., 2007, Medeiros, 2015).

Unfortunately, transforming data into high-quality biomarkers that are robust and predictive is challenging (Wang et al., 2017). In this Chapter we present an overview of different approaches how to perform biomarker discovery in the context of medical imaging. The list of strategies is not intended to be complete, rather we focus on how recent developments in machine learning can facilitate biomarker discovery.

4.1 Supervised prediction

The traditional approach of biomarker discovery is illustrated in Figure 4.2(a). It involves the development of a hypothesis based on professional experience, a specific observation or theoretical motivation. Studies are then conducted to verify or reject this hypothesis. Traditional machine learning methods are trained with hand-crafted features on specific manually defined outcome variables. The investigated features (which are the biomarker candidates) are defined manually, a priori. On one hand, this allows a more direct interpretation of the prediction results since the used features are known, e.g. using the feature importance measures provided by random forests (Breiman, 2001). On the other hand, this limits the data exploration to manually defined candidates. Moreover, outcome variables are also needed to develop and train the model.

There are various examples for this type of approach. The capability for early breast cancer diagnosis of manually defined serum features was evaluated in Opstal-van Winden et al. (2012), where antigen protein concentrations in serum samples were used to train a

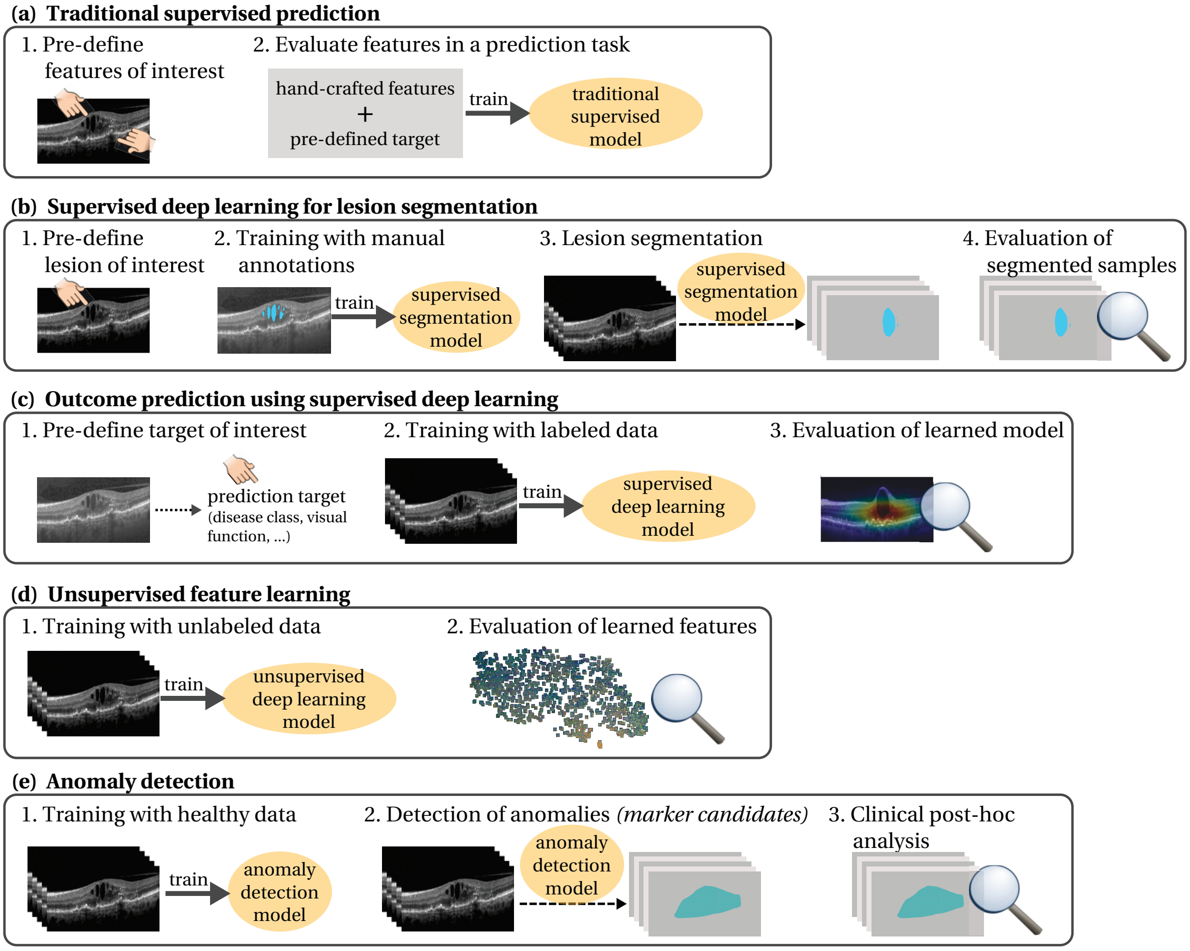

(a) Traditional supervised prediction
1. Pre-define features of interest
2. Evaluate features in a prediction task
hand-crafted features
+
pre-defined target
train
traditional supervised model
(b) Supervised deep learning for lesion segmentation
1. Pre-define lesion of interest
2. Training with manual annotations
train
supervised segmentation model
3. Lesion segmentation
supervised segmentation model
4. Evaluation of segmented samples
(c) Outcome prediction using supervised deep learning
1. Pre-define target of interest
prediction target
(disease class, visual function, ...)
2. Training with labeled data
train
supervised deep learning model
3. Evaluation of learned model
(d) Unsupervised feature learning
1. Training with unlabeled data
train
unsupervised deep learning model
2. Evaluation of learned features
(e) Anomaly detection
1. Training with healthy data
train
anomaly detection model
2. Detection of anomalies (marker candidates)
anomaly detection model
3. Clinical post-hoc analysis

Figure 4.2: Various strategies for biomarker discovery are depicted. While the first two strategies (a-b) investigate pre-defined lesions and can be seen as hypothesis verification procedure, the other shown approaches (c-e) can be interpreted rather as hypothesis-generation strategies.

random forest model. In Lad et al. (2018), the thickness of inner retinal layers was evaluated as biomarker in the context of Alzheimer's disease. Thickness measurements of specific retinal layers were also investigated in Farsiu et al. (2014), namely to predict the presence of AMD with a linear model (Dobson and Barnett, 1990).

4.2 Supervised deep learning for lesion segmentation

An alternative strategy involves the segmentation of specific marker candidates in images (Figure 4.2(b)). Supervised deep learning offers a powerful alternative to traditional machine learning algorithms. While non-data driven approaches require domain expert knowledge to perform careful engineering of hand-crafted features for model training, supervised deep learning methods avoid biases due to manual design of features by automatically learning from data.

Based on a hypothesis, first a specific image structure of interest is defined. Then, manual annotations need to be generated by experts in order to train a deep learning algorithm to segment these structures of interest. The trained segmentation model can then be used to segment large-scale image datasets, which allows a subsequent statistical analysis of the segmented structure. Furthermore, the segmentation results can also be used as features in the traditional pipeline of biomarker discovery (Section 4.1).

This strategy utilizes the power of deep learning approaches and allows a large-scale analysis. However, this analysis is limited to a priori defined markers of interest, requires manual annotation of images which can be costly to obtain and the results may suffer from intra- or inter-grader variability (Asman and Landman, 2011). Moreover, the training set should consist of a representative patient cohort in order to achieve an acceptable generalization performance.

Supervised lesion segmentation is presented in Schlegl et al. (2018), where a deep learning algorithm is trained to segment macular fluid in OCT scans. The task of segmenting brain lesions is tackled in Kamnitsas et al. (2017), Nair et al. (2018), Roy et al. (2018). An example for using segmentation results in a subsequent analysis is this is the work of Schmidt-Erfurth et al. (2018a), where automatically segmented OCT structures such as IRF and SRF were correlated with visual function (BCVA) and used to predict BCVA at a future time point.

4.3 Outcome prediction using supervised deep learning

The above mentioned strategies are effective for checking a specific hypothesis or investigating a priori defined biomarker candidates. However, they are limited to pre-defined features or structures of interest. Alternatively, supervised deep learning models can be trained on large-scale annotated data to predict particular outcome variables, including diagnoses or functional parameters. Assuming that the networks will learn to capture features in the input images that are relevant for the given task, a subsequent analysis of the model is performed (Figure 4.2(c)). Appropriate visualization techniques are needed to identify the learned features which are important for the prediction task and to understand the properties of the model. Though this evaluation allows the exploration of relationships between the target value and the input data, it represents a difficult and complex task that needs expert knowledge. In addition, due to the complexity of the prediction task, a large number of samples with known or annotated outcomes which form a representative patient cohort is needed to train the model. Moreover, only features that are relevant for the defined target variable are learned, which means that features that are relevant for other tasks or different problem definitions might be missed.

In Rajpurkar et al. (2017), a supervised deep learning model was trained to detect pneumonia based on chest X-rays as input. They used a technique called class activation mapping (CAM) (Zhou et al., 2016) to interpret the network predictions by producing heatmaps that visualize which areas of the input image were most relevant for the prediction. Cardiovascular risk factors were predicted from fundus photographs in Poplin et al. (2018). Here, the soft attention heatmap approach (Xu et al., 2015) was used to identify the anatomical regions that were relevant for the model predictions. Esteva et al. (2017) presented a deep learning method to classify skin cancer using images and provided saliency maps for a qualitative evaluation. Another visualization technique called occlusion testing was used in Kermany et al. (2018), where retinal diseases are classified based on OCTs.

4.4 Unsupervised feature learning

The main drawback of the supervised strategy presented in Section 4.3 is the requirement of large-scale annotated data. For instance, Rajpurkar et al. (2017) used 112,120 images to train the model, the training in Esteva et al. (2017) involved around 130,000 images and Poplin et al. (2018) used samples coming from over 284,000 patients. Clearly, this amount of labels is costly, and infeasible to obtain in many scenarios. Unsupervised deep representation learning on unlabeled datasets offers an alternative approach to automatically learn features from the data, omitting the need for large-scale annotations. The idea is to capture phenotype characteristics of the patient population with the learned features (Figure 4.2(d)). The assumption is that they will be disease-specific but general at the same time. General in the sense that they are capturing prominent characteristics of the patient population, and not only features which are important for a particular task (cf. Section 4.3). In other words, this means that the learned features are not biased towards a specific prediction task. At the same time, this approach may be prone to ”overlook” characteristics which appear rarely (but are nevertheless important for a specific task). The learned features can be evaluated in various ways, e.g. the discriminative power can be assessed in a prediction task by training a separate classifier with the learned features. As for the previously presented strategies, a representative patient cohort is needed to train the model here as well. Chapter 7 presents a method that performs unsupervised representation learning from images with the aim to identify relevant disease-specific phenotype characteristics.

4.5 Anomaly detection

Anomaly detection offers an interesting alternative in the context of biomarker discovery. It is defined as the family of approaches that detect samples which differ from the distribution of normal data used during training (Pimentel et al., 2014). In general, anomaly detection can be seen as a two-step process: first a model of normal appearance is learned, and then it is applied to detect deviations from this normal data (anomalies) during test time (Pimentel et al., 2014). Anomaly detection methods have been proposed in various disciplines such as hyperspectral remotely sensed imagery (Matteoli et al., 2014), video surveillance (Del Giorno et al., 2016, Kumaran et al., 2019), in medicine (Baur et al., 2018, Schlegl et al., 2017), or others (Chalapathy et al., 2018, Erfani et al., 2016). For instance, Erfani et al. (2016) presented a hybrid approach where a representation is learned by a neural network in a first stage and a One-Class SVM is used to detect anomalies by estimating the distribution of healthy samples. Inspired by this idea, other works have developed one class neural networks (OC-NNs), combining both stages into an end-to-end model to learn features that are explicitly optimized towards the one-class objective (Chalapathy et al., 2018, Ruff et al., 2018). Another group of techniques utilizes autoencoders to detect anomalies, assuming that the model fails to reconstruct out-of-distribution samples not observed in the training set, leading to a high residual error for anomalies (Ribeiro et al., 2018, Zhou and Paffenroth, 2017). Schlegl et al. (2017) proposed an unsupervised anomaly detection based on generative adversarial networks (GANs) in retinal OCT images. The model (*AnoGAN*) is trained on patches coming from healthy eyes to learn the variability of normal images, and detects abnormal patches in new data by comparing their fit to the learned distribution. An extension of this approach has been proposed to overcome the issue of computational inefficiency during detection time (Schlegl et al., 2019). Other works that use GAN based methods for anomaly detection have been presented for telecom fraud detection (Zheng et al., 2018), abnormal crowd behavior in videos (Ravanbakhsh et al., 2017) as well as for network activity and natural images (Zenati et al., 2018). Furthermore, long short term memory (LSTM) networks have been used for anomaly detection in sequential data (Ergen et al., 2017, Malhotra et al., 2015).

Generally, the detection can be performed on different levels of detail (e.g. on image or pixel level), where the detected anomalies are either known disease markers or new *biomarker candidates*. While screening of clinical routine data to identify already known lesions is one possible use-case, anomaly detection can also constitute a first step in the process of biomarker discovery. In the latter case, the identified anomalies can be subsequently transformed from marker candidates to effective markers in a clinical post-hoc analysis, establishing them as novel biomarkers (Figure 4.2(e)).

Anomaly detection methods are trained only on normal data without labels. This omits the need of collecting a representative patient cohort with an appropriate amount and variations of pathologies for training. While capturing all possible disease manifestations or disease related characteristics is a challenging task (e.g. rare diseases), collecting a healthy set of samples that captures the normal appearance can be less difficult (Markou and Singh, 2003). In addition, no annotations are needed, since the targets of interest (anomalies) are implicitly defined by the appearance of the normal training set. This means that this approach is not limited to a specific disease or marker category, considering the fact that the training does not rely on explicit a priori descriptions of markers and all possible deviations from normal are detected by definition.

Both Chapter 5 and Chapter 6 present anomaly detection approaches for retinal OCT images.

CHAPTER 5

Unsupervised Identification of Disease Markers in Retinal OCT Imaging Data

"Those who love wisdom
must investigate many things."

– Heraclitus

This Chapter contains the first manuscript *"Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data"* of the thesis, which has been published in the journal *"Transactions on Medical Imaging"*. In this manuscript, we propose to identify and categorize anomalies as marker candidates in retinal OCT images in an unsupervised way. In a first step, an anomaly detection approach is used to segment anomalous regions on pixel-level. A multi-scale deep denoising autoencoder is trained on healthy scans, and a One-class SVM (Schölkopf et al., 2001) is used to estimate the distribution of normal appearance based on the learned feature representation. The trained system can then be applied to segment anomalies in new data. In a second step, clustering in the anomalies identifies stable categories. We compare our anomaly detection approach to alternative unsupervised feature learning techniques and perform both qualitative and quantitative evaluation to assess the performance of the proposed method.

Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data

Philipp Seeböck, Sebastian M. Waldstein*, Sophie Klimscha, Hrvoje Bogunovic, Thomas Schlegl,
Bianca S. Gerendas, René Donner, Ursula Schmidt-Erfurth, and Georg Langs

***Abstract*—The identification and quantification of markers in medical images is critical for diagnosis, prognosis, and disease management. Supervised machine learning enables the detection and exploitation of findings that are known *a priori* after annotation of training examples by experts. However, supervision does not scale well, due to the amount of necessary training examples, and the limitation of the marker vocabulary to known entities. In this proof-of-concept study, we propose unsupervised identification of anomalies as candidates for markers in retinal Optical Coherence Tomography (OCT) imaging data without a constraint to a priori definitions. We identify and categorize marker candidates occurring frequently in the data, and demonstrate that these markers show predictive value in the task of detecting disease. A careful qualitative analysis of the identified data driven markers reveals how their quantifiable occurrence aligns with our current understanding of disease course, in early- and late age-related macular degeneration (AMD) patients. A multi-scale deep denoising autoencoder is trained on healthy images, and a one-class support vector machine identifies anomalies in new data. Clustering in the anomalies identifies stable categories. Using these markers to classify healthy-, early AMD- and late AMD cases yields an accuracy of 81.40%. In a second binary classification experiment on a publicly available data set (healthy vs. intermediate AMD) the model achieves an AUC of 0.944.**

***Index Terms*—unsupervised deep learning, anomaly detection, biomarker identification, optical coherence tomography**

I. Introduction

The detection of diagnostically relevant markers in imaging data is critical in medical research and practice. Biomarkers are required to group patients into clinically meaningful subgroups regarding disease, disease progression, or treatment response. Imaging data provides a wealth of information relevant for this grouping in the form of *imaging biomarkers*. Typically, image analysis methods are trained based on *a priori* defined categories, and annotated imaging data. This makes large-scale annotation necessary, which may be costly or infeasible, limits detection to known marker categories, and, overall, slows down the discovery of novel markers. In contrast, unsupervised detection of anomalies and subsequent data-driven identification of new markers offer the possibility for unbiased classification of a disease and the identification of novel risk factors. Unsupervised detection can extend our knowledge about the underlying pathophysiology of diseases. The resulting biomarkers can enable a description of the entire spectrum of a disease, from the earliest manifestations to the terminal stages [1]. In this proof-of-concept study, we perform anomaly detection on retinal images to identify biomarker candidates, categorize them, and evaluate their link to disease.

A. Clinical background

OCT [2] provides high-resolution, 3D volumes of the retina and is the most important diagnostic modality in ophthalmology. Approximately 30 million ophthalmic OCT procedures are conducted per year worldwide, on par with imaging modalities such as magnetic resonance imaging, computed tomography, and positron emission tomography [3]. Each position of the retina sampled by an optical beam results in a vector, the A-scan. Adjacent A-scans form a 2D slice, alias B-scan, which consecutively form the entire volume. Examples of B-scans are shown in Fig. 2 on the left.

Retinal diseases causing vision loss affect many patients. For instance, age-related macular degeneration (AMD) is the leading cause of blindness in industrialized countries and has a worldwide prevalence of 9% [4]. Even-though intraretinal fluid shows some predictive value [5], we are lacking accurate and reliable imaging markers and predictors for individual patients disease courses. The discovery of novel reliable markers in imaging data is relevant to enhance individual care, encompassing the identification and categorization of marker candidates, and the quantification of their link to disease. Not all patterns occurring in OCT volumes are understood or interpretable, and for certain retinal diseases such as for AMD [6], pathogenic mechanisms are not yet fully known.

Computational anomaly detection [7] and categorization is a natural approach to tackle this problem, where the former is defined as the detection of cases that differ from the normal samples available during training. In retinal images this is a difficult task for many reasons. In contrast to natural

Manuscript received November,29,2017; revised August,08,2018; accepted October, 13, 2018.

The financial support by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development, by the Austrian Science Fund (FWF I2714-B31), and by IBM (2016-2017 IBM PhD Fellowship Award, and Faculty Award) is gratefully acknowledged. A Tesla K40 used for this research was donated by the NVIDIA Corporation.

P. Seeböck, T. Schlegl, R. Donner and G. Langs are with the Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Austria (email: philipp.seeboeck@meduniwien.ac.at, georg.langs@meduniwien.ac.at)

P. Seeböck, S.M. Waldstein, H. Bogunovic, S. Klimscha, B.S. Gerendas, U. Schmidt-Erfurth and G. Langs are with the Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology and Optometry, Medical University Vienna, Austria. (email: sebastian.waldstein@meduniwien.ac.at)

* corresponding author

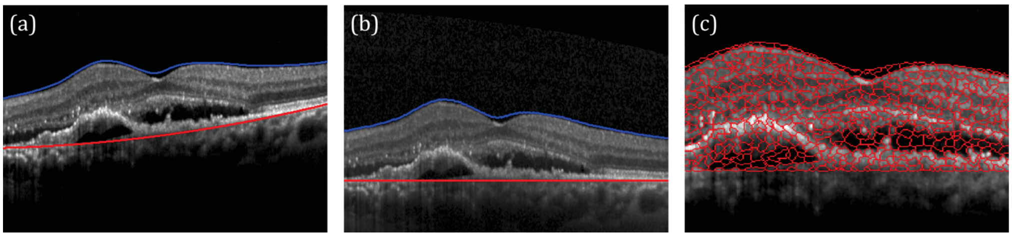


Fig. 1. Preprocessing of OCTs. (a) The original OCT scan with the top layer highlighted in blue and the bottom layer (Bruch's membrane) in red. (b) The same scan after normalization was applied (shift to horizontal plane, brightness, contrast). (c) A zoomed-in snippet of the over-segmentation result.

images, such as photographs, in retinal imaging the findings relevant for diagnosis cover only a small fraction of the overall volume. Furthermore, their deviation from normal tissue is subtle compared to the variability of healthy retinas. Therefore, to identify novel marker candidates, we form a model of normal tissue variability, and detect anomalies deviating from this model. In this paper, we define *normal* as the absence of pathological changes beyond age-related alterations, i.e. the only allowed visible alteration included drusen below 63 μm in size according to the Beckman Initiative Classification [8]. This definition accounts for age-related changes that normally do not result in visual impairment. For instance, the majority of elderly patients shows small drusen, while still maintaining normal vision.

Some retinal diseases such as retinal vein occlusion (RVO), often occur unilaterally. Thus, the contralateral eye is not affected by the acute event of the disease and can be elegantly used as training data for the normal appearance model. In our case, these volumes of contralateral eyes were screened by a retinal specialist to rule out cases with pathological changes. Since our model is purely trained on normal data, we omit the need to collect a dataset containing a sufficient amount of anomalies representing the entirety of their possible variability. At the same time, the applicability of the model is not limited to a specific disease.

B. Related Work

Anomaly detection can be a crucial first step in the process of biomarker detection. The results of these algorithms are affected by the quality of the features used for characterizing the data. Supervised deep learning has recently improved the state-of-the-art in various tasks, such as image classification [9], object detection [10] or weakly supervised learning linking semantic descriptions to image content [11]. It results in rich feature representations, although at the cost of requiring large amounts of annotated training samples and the limitation to known markers. On the other hand, unsupervised learning enables the exploitation of unlabeled data, capturing the structure of its underlying distribution [12]–[15]. A well-known and widely used technique for feature learning is Principal Component Analysis (PCA) [16], which is computationally efficient, but limited to a linear embedding. In contrast, unsupervised deep learning of convolutional neural networks (CNNs) is computationally more expensive, but can learn a non-linear embedding.

In [13], unsupervised CNN training was performed by discriminating between surrogate image classes created by manually defined data augmentation to render the resulting representation robust to certain transformations. Other studies propose incorporation of supervisory signals such as spatial context [12] to omit the requirement of manually annotated data. In our study we identify clinically relevant biomarkers without prior human input, which could bias the result.

The proposed anomaly detection method is inspired by the combination of Deep Belief Networks (DBNs) with One-Class SVMs for anomaly detection in real-life datasets, which have considerably different characteristics compared to medical images [17]. The DBNs learn a feature representation, while the One-Class SVM finds a boundary describing regions in the feature space with high probability density of the training data. Erfani *et al.* [17] trained DBNs in a layer-wise fashion, and did not use a multi-scale architecture, as we did. According to [14], [15], the combination of joint training of layers and local regularization constraints for each layer is more advantageous than layer-wise training without constraints. Therefore, we both trained the deep convolutional autoencoder (DCAE) [14] and the deep denoising autoencoder (DDAE) [15] jointly. The multi-scale architecture was partly inspired by [11], where weakly supervised learning was used to link image information to semantic descriptions of image content. Since DCAEs are specifically designed to learn effective representation of images, they are a logical comparison method when learning unsupervised image representations. The idea of using normal subjects to model a normal population is not novel. Sibide *et al.* [18] modeled the appearance of normal OCT B-scans with a Gaussian Mixture Model (GMM) and detected anomalous B-scans as outliers. The number of outliers served as basis for classification of an entire OCT volume. In contrast, we aim at pixel level anomaly detection. In [19], a shape model of normal retinal layers is used to segment anomalies. The model has a close fit in normal regions, while there is no fit in areas of anomalous shapes. The limitations of this approach are that it heavily depends on the quality of the layer segmentation algorithm and does not take into account image information explicitly to detect anomalies. Finally, Schlegl *et al.* [20] proposed AnoGAN, a deep convolutional Generative Adversarial Network (GAN) to learn a manifold of normal anatomical variability, in order to identify anomalous regions in OCT images. AnoGAN is restricted to healthy representations by definition, which makes it inappropriate for a straightforward subsequent clustering step of anomalies. In contrast, we aim at learning a feature representation with our autoencoder approach which is general enough to enable a meaningful embedding of anomalies, though we solely need normal training data.

Examples for the classic biomarker identification strategy are [21], [22], where the authors used a priori defined features in a supervised way to evaluate the applicability as biomarkers for specific diseases. In contrary, our approach focuses on identifying new marker candidates in an unsupervised way.

Regarding classification of retinal diseases on volume level, main focus of related work [21], [23] is to solve the classification task itself. In contrast, here our target is to evaluate the link of identified categories, alias marker candidates, to disease by using them as features for classification.

C. Contribution

We propose a method to identify marker candidates in imaging data in an unsupervised fashion. Our approach first separates anomalous candidates from normal tissue in retinal spectral-domain Optical Coherence Tomography (SD-OCT) based on the features learned by *DDAE* on healthy samples, and a One-Class SVM to model normal appearance distribution. We identify categories of frequently occurring anomalies using clustering, and evaluate their link to disease. In a qualitative evaluation, retinal experts could map part of the categories identified by this approach to known retinal structures. At the same time other categories remain as novel anomaly candidates, for which results on the classification tasks suggest that they are also linked to disease.

This paper is an extension of our previous work [24] introducing a new feature-learning approach, and more in-depth evaluation of anomaly detection, categorization, and the link of these marker candidates to disease.

II. Methods

To capture visual information at different levels of detail, we used a multi-scale approach to perform superpixel-wise segmentation of the visual input. While the preprocessing steps are shown in Fig. 1, the overall architecture is illustrated in Fig. 2. After preprocessing (Section II-A), 2D-patches extracted from B-Scans from healthy OCT volumes were used to train a deep denoising autoencoder model (Section II-B), which provided an embedding that represented healthy anatomical variability. A One-Class SVM was trained on this embedding to obtain a boundary, which encompassed the distribution of healthy patches (Section II-C). Using this boundary, unseen volumes (i.e. volumes not used during training) were segmented into healthy or anomalous regions. Subsequent clustering of anomalous regions partitioned anomalies into more specific categories (Section II-D).

A. OCT Preprocessing

For all OCT volumes, we identified the top (Internal Limiting Membrane - ILM) and bottom (Bruch's Membrane - BM) layer of the retina using a graph-based surface segmentation algorithm [25], where the bottom layer was used to flatten the retina by projecting it to a horizontal plane, as shown in Fig. 1(b). The top and bottom layer of the retina are also illustrated as blue and red in Fig. 2 at the bottom left. This reduced the differences in appearance caused by varying orientations and positions of the retina within the volume. We applied brightness and contrast normalization for each B-scan and added a constant to shift the values into a positive range. The latter was necessary to ensure that the deep denoising autoencoders ($DDAE_1$, $DDAE_2$) were able to reconstruct the input patches ($\dot{x}$, $\ddot{x}$) properly during training. Finally, we performed over-segmentation of B-scans to monoSLIC superpixels, s, of an average size of 4×4 pixels [26], as illustrated in Fig. 1(c). This merges pixels into homogeneous groups of superpixels, which allows to perform the computations on a reduced number of superpixels as opposed to computations on every pixel.

Preprocessing of healthy B-scans, I_h, with $h = 1, \ldots, H$, resulted in S^h superpixels, s_i^h, for each (as illustrated in Fig. 1), with center positions p_i^h and $i = 1, \ldots, S^h$, where H denotes the number of healthy B-Scans, S^h the number of superpixels per B-Scan, i the index of the superpixel, and h the index of the healthy B-Scan.

B. Unsupervised Learning of Appearance Descriptors

The network architecture of the deep denoising autoencoder consists of an encoding and decoding part. We chose three fully connected layers, with 2048 neurons in the first, 1024 in the second, and 512 in the third layer to build the encoder, with the structure also denoted as `2048f-1024f-512f`. The mirrored version of the encoder (`512f-1024f-2048f`) formed the decoder, as illustrated in Fig. 2. The weight matrices of two corresponding layers were tied: $W_{enc} = W_{dec}^T$. All layers were followed by Exponential Linear Units (ELUs) [27], with $\alpha = 1$:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(exp(x) - 1) & \text{if } x \leq 0 \end{cases} \tag{1}$$

The Mean Squared Error function, $MSE(x, \hat{x})$, was chosen as a loss function for training, where x denotes the input patch and $\hat{x}$ the output of the last layer of the decoder. The autoencoder was trained jointly in an end-to-end fashion, as proposed in [15]. In addition, we added a local constraint to each layer by corrupting the input of every layer in the encoder. More precisely, a fraction of the inputs was set to zero. As opposed to layer-wise training, this corresponds to unsupervised joint training with local constraints in each layer.

We conducted unsupervised training of two deep denoising autoencoders ($DDAE_1$, $DDAE_2$) on the patches, $\dot{x}_i^h$ and $\ddot{x}_i^h$, extracted at center positions of superpixels p_i^h from the healthy B-scans, I_h. While $\dot{x}_i^h = 32 \times 32$ served as input for $DDAE_1$, $DDAE_2$ was trained with 128×32 patches $\ddot{x}_i^h$, downsampled to 32×32. The provided patch sizes are given in pixels. Both models were fixed for the subsequent training of another denoising autencoder, $DDAE_3$, its single-layer architecture denoted as `256f`, with the concatenated feature vectors $[\dot{y}\ddot{y}]$ as input, where $\dot{y}_i^h = DDAE_1(\dot{x}_i^h)$ and $\ddot{y}_i^h = DDAE_2(\ddot{x}_i^h)$. All three learned encoders from

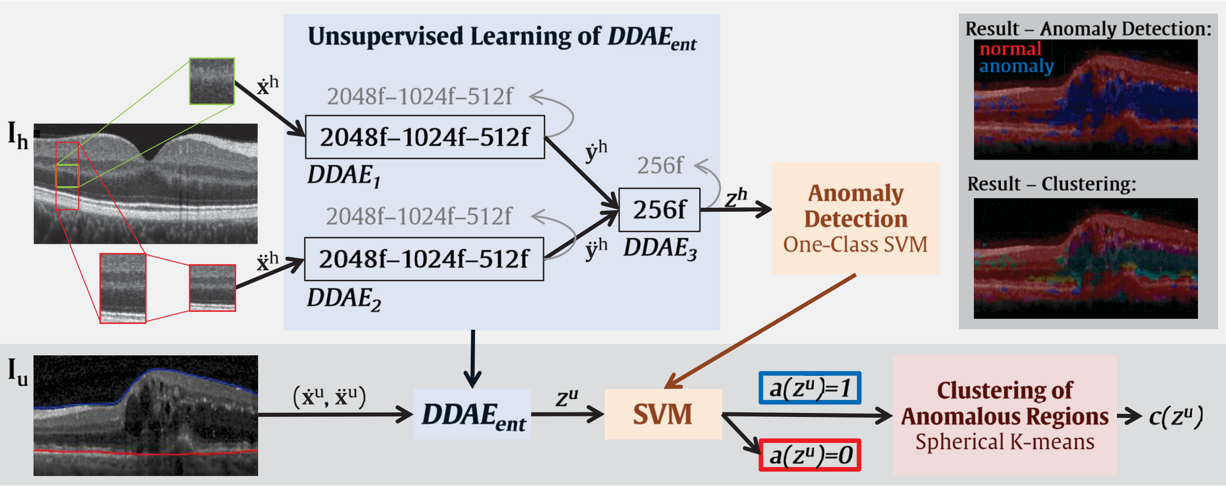


Fig. 2. Multi-scale architecture used in the experiments. Pairs of $\dot{x}_i^h = 32 \times 32$ (green) and 128×32 (red) patches were extracted at positions p_i^h in healthy OCTs I_h from B-scans, where the larger patches were downsampled to $\ddot{x}_i^h = 32 \times 32$, as illustrated on the left. The encoders of all unsupervised learning modules are denoted in black, while the decoders are depicted in gray. Once learned, the encoders of $DDAE_{ent}$ were used to create the new feature representation z_i^h, for a specific superpixel s_i^h, with position p_i^h in the B-scan. Subsequent training of a One-Class SVM enabled the opportunity to detect anomalous regions, $a(s_i^u) = 1$, in unseen B-scans I_u (i.e. B-Scans not used during training). In order to subdivide anomalous regions into meaningful categories, c_j, clustering was performed. This means that for each superpixel s_i^u with position p_i^u in an anomalous region, a cluster assignment $c(s_i^u)$ was performed. An example of anomaly detection and subsequent clustering of anomalous regions is shown on the top right.

$DDAE_1$, $DDAE_2$, and $DDAE_3$ together formed the final model, $DDAE_{ent}$, that gave us a 256 dimensional feature representation, $z_i^h = DDAE_{ent}(\dot{x}_i^h, \ddot{x}_i^h) = DDAE_{ent}(s_i^h)$, for a specific superpixel, s_i^h, with corresponding patches $(\dot{x}_i^h, \ddot{x}_i^h)$ extracted at the central position of the superpixel, p_i^h. The multi-scale architecture allows to incorporate the local information of the smaller patch and at the same time the neighborhood and orientation information of the larger patch.

C. Anomaly Detection with One-Class SVM

Based on the learned feature representation, z_i^h, we estimated the distribution of healthy examples with a One-Class SVM [28], using a linear kernel. The SVM searches for a boundary that describes the distribution of normal data, which serves as a decision boundary for unseen data. New samples can then be classified either as coming from the same data distribution if lying inside the boundary (0, normal) or not (1, anomaly). Since we used a linear kernel for One-Class SVM, the only hyper-parameter was ν. This parameter determines the amount of normal training data that must lie within the boundary, i.e., which is detected as normal. For example, a value of 0.1 means that 90% of the training samples are within the boundary. In this work, we chose the parameter value with the highest dice score on the validation set for the final model.

For unseen B-Scans, I_u, with $u = 1, \ldots, U$, features z_i^u and the corresponding class $a(z_i^u) = \{0, 1\}$ were computed for each superpixel, s_i^u, with position p_i^u within the top and bottom layer of the retina, where U denotes the number of unseen B-Scans. The computed class label $a(z_i^u)$ was assigned to the entire superpixel: $a(z_i^u) = a(s_i^u)$. This provided a segmentation of the retina into two classes at superpixel level.

D. Categorization of Anomalous Regions

We used spherical K-means clustering [29] with cosine distance to sub-segment anomalous superpixels $a(s_i^u) = 1$, which have been classified as anomalous by our method in the first stage, in unseen B-Scans into C clusters, $c(s_i^u) = c(z_i^u) = j$, with $j = 1, \ldots, C$. More precisely, we trained a cluster model using the 256 dimensional feature representation z on an ”anomaly training set”, that was composed of samples with $a(z_i^u) = 1$ only, to obtain cluster centroids c_j. The number of cluster centroids, C, was determined by an internal evaluation criterion called the Davies-Bouldin (DB) index [30], calculated on the anomaly training set. A small value indicates compact and well-separated clusters, hence, the model with the smallest DB index was selected.

To segment an unseen B-Scan I_u, each superpixel with the property $a(s_i^u) = 1$ got a cluster assignment, $c(s_i^u)$, where $c(s_i^u)$ gives the index, j, of the nearest cluster centroid, c_j. To facilitate reading, we omitted indices i, h, and u henceforth.

III. Evaluation

Our evaluation tests: (1) if the proposed algorithm can identify anomalous regions in imaging data, (2) if the algorithm can detect stable categories of anomalies, and (3) if these categories can serve as disease markers.

Data: We used n=786 OCT volumes from just as many patients from our database, which was divided into six subsets[1]: *Healthy* (*training* n=283, *test* n=33), *late AMD* (*categorization* n=362, *validation* n=5, *test* n=26), and *early AMD categorization* n=77. The volumes of *healthy training* and *test* were selected from 482 / 209 contralateral eye scans of patients with RVO / AMD in the other eye. Volumes with pathological changes beyond age-related alterations were excluded. The

[1] An overview of the data and experiments can be found in the supplementary material

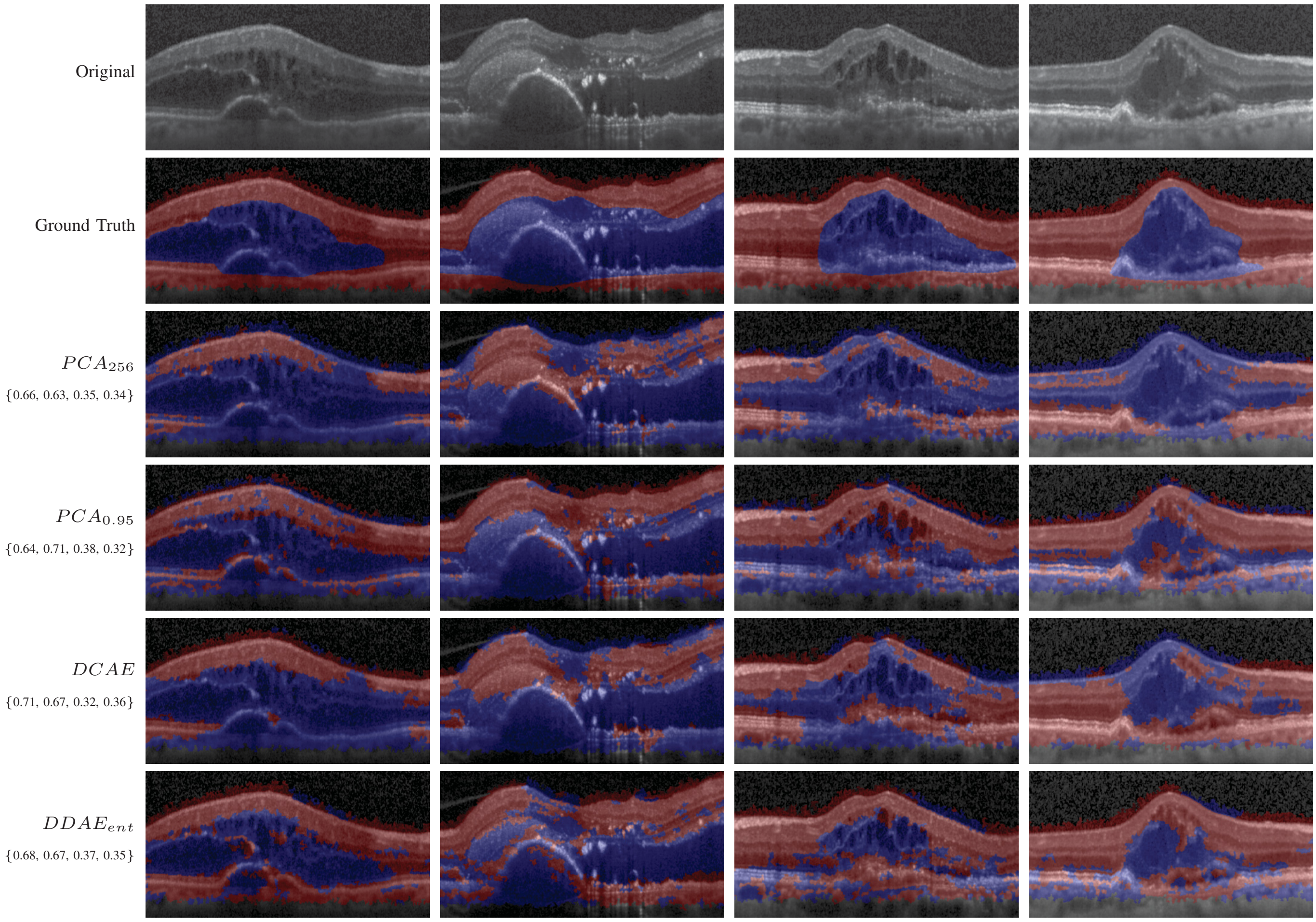


Fig. 3. Anomaly detection: for each column, the same B-scan is illustrated, where red and blue colors indicate healthy and anomalous areas. To guarantee objectivity, the first (last) two columns show examples with highest (lowest) dice of $DDAE_{ent}$. Dice values are provided for each method and sample. The full quantitative evaluation result is given in Table I.

volumes of *late AMD* were eyes with active neovascular AMD, where a retina specialist manually annotated all areas that contained pathologic features in 31 OCT volumes. These volumes with voxel-wise annotations of anomalous regions were randomly divided into *late AMD validation* and *late AMD test*. The volumes in *early AMD categorization* were classified by clinical retina experts as early, non-neovascular AMD according to [8]. All image data was anonymized and ethics approval was obtained for the conduct of the study from the ethics committee at the Medical University of Vienna.

The volumes were acquired using Spectralis OCT instruments (Heidelberg Engineering, GER), with a voxel dimensionality of $512 \times 496 \times 49$, which depicted a $6mm \times 2mm \times 6mm$ volume of the retina, with the voxel spacing of $11\mu m \times 4\mu m \times 120\mu m$. Thus, one OCT volume is composed of 49 B-scans, where the distance between B-scans is $120\mu m$. All volumes were preprocessed as described in Section II-A. Due to the anisotropy of the imaging data, the proposed approach works with 2D patches extracted from B-scans. Pairs of image patches with pixel size of 32×32 and 128×32 were extracted, illustrated for a single position in Fig. 2, on the left.

Additionally, we used 384 Bioptigen SD-OCT volumes (269 intermediate AMD, 115 control) from a publicly available dataset [21]. Since this dataset differs in appearance from our database (different OCT vendor), we conducted additional preprocessing steps: non-local means noise filtering, resizing B-scans to match the Spectralis B-Scans in resolution, and adjustment of image intensity values. Details can be found in the supplementary material.

Training Details: All networks were trained on the healthy training set for 300 epochs and used tied weights. A validation set of five OCT volumes was used for parameter tuning. Due to limited computational resources, only a small parameter selection was assessed. We used standard values for ELU ($\alpha = 1$), momentum (0.9), and mini-batch (50). The initial learning rate was set to the highest value that did not diverge (0.0001) for 150 epochs, and decreased to 0.00001 for another 150 epochs. We experimented with two different corruption values (0.5, 0.9) for fully connected layers, and we found 0.5 to work better. We also conducted experiments with shallower network architectures, which we empirically found to work slightly worse. Since the One-Class SVM hyper-parameter ν is bounded between 0 and 1, we varied ν between 0.01 and 0.9 for all methods: $\nu = [0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. For the experiments we used the *Torch7* framework [31] and the One-Class

SVM implementation of *libsvm* [32].

A. Evaluation of Anomaly Detection

The anomaly detection model was trained on the *healthy training* set, with 277340 extracted pairs of image patches, randomly selected from all 13867 B-Scans between the top and bottom layer of the retina to avoid retrieving numerous "background patches" without relevant content. While the influence of ν was analyzed on *late AMD validation* volumes, the final performance of the learned anomaly detection model was evaluated on *late AMD test*. Each B-scan of all validation and test volumes was expert annotated between the ILM and BM layer, and *dice*, *precision*, *recall*, *specificity* and *accuracy* were calculated for algorithmically detected anomalous regions with regard to the manual annotations on volume level.

We compared the $DDAE_{ent}$ model to the feature learning approaches described in Section I-B: PCA embedding, and a Deep Convolutional Autoencoder, denoted as $DCAE$. PCA was chosen since it is a well-known and widely used technique for feature learning. At the same time we were aiming at a powerful representation of images, which is why DCAEs, that are specifically designed for images, are a logical comparison method [14]. To ensure a fair comparison for PCA, we trained two models. In the first model, PCA_{256}, the dimensionality was chosen to match the feature dimension, z, of the proposed model. For both scales, the first 128 principal components were kept. In the second model, $PCA_{0.95}$, for each scale, the first components that described 95% of the variance were kept. To retrieve the final feature representation, the feature vectors $\dot{y}$ and $\ddot{y}$ of both scales were concatenated to obtain z.

For $DCAE$, we use an encoder with a convolutional layer (`c`) and 512 9×9 filters, followed by a 3×3 non-overlapping max pooling (`p`) and two fully connected layers with 2048 and 512 units (`512c9-3p-2048f-512f`). The decoder was composed of deconvolution (`dc`) and unpooling (`up`) layers to approximately invert the output of the encoder and reproduce the input (`512f-2048f-3up-512dc9`). All layers, except pooling and unpooling, were followed by ELUs. This DCAE-architecture replaced $DDAE_1$ and $DDAE_2$, while the architecture of the third model($DDAE_3$) remained the same. To ensure a fair comparison, the feature dimension of the individual model outputs matched the dimensionality of the proposed method.

B. Evaluation of Anomaly Categorization

We trained two clustering models on two different datasets: *late AMD categorization* and *early AMD categorization*. We extracted 354760 and 75460 pairs of image patches, respectively. For both models, we varied the number of clusters, C, between 2 and 30 and selected the clustering model with lowest DB-index. In order to qualitatively evaluate the categories found in the regions identified as anomalous, a segmentation of the retina based on the identified anomalous categories was computed on both datasets. Assignment of each pixel was based on the learned centroids and the nearest-cluster-center labeling (Section II-D). The used manual annotations of the test set provided only a binary distinction into healthy and anomalous, and did not describe all anomalies that were visible in separate categories. Therefore, two clinical retina experts conducted a qualitative evaluation of the results by visually inspecting the results. While the number of clusters was determined by the DB-index, category descriptions were identified by the experts. Additionally, cosine distances between centroids of the two cluster models trained on *late AMD categorization* and *early AMD categorization* were computed in order to evaluate the correspondence between both models.

C. Evaluation of Volume Level Disease Classification

To evaluate if the identified categories can serve as disease markers and encode valuable discriminative information, we used the segmentation of the retina into C clusters, originating from the clustering model learned on *late AMD categorization*, to conduct multi-class classification on patient level. The volume of each cluster served as feature vector for every case. Since the clusters that were identified in *early AMD categorization* could all be mapped to clusters identified in *late AMD categorization*, which at the same time revealed one additional cluster, we used only the latter more complete anomaly category set as basis for these experiments. We trained a random forest (RF) classifier [33] (#trees=64) with these feature vectors using a set of randomly chosen late AMD, early AMD and healthy cases (n=50 per class) from the training and categorization sets. We then applied the classifier to a separate test set not involved in anomaly detection, categorization, or classifier training composed of *late AMD test* (n=26) the remaining part of *early AMD categorization* (n=27) and *healthy test* (n=33). We report feature importance values obtained by random forest training, and the classification accuracy on the test set.

For comparison, we trained a second RF model without category information. Using the same evaluation setting as described above, we used the binary segmentation of the retina as features, originating from anomaly detection (Section II-C), instead of the learned clusters.

We performed a second experiment to evaluate how the method generalizes to a dataset of a different vendor. The Bioptigen volumes were used for a second volume classification experiment. Following the evaluation procedure in [23], the dataset was randomly divided into *bioptigen training* (218 AMD, 65 control) and *bioptigen test* (50 AMD, 50 control), and the RF was trained with #trees=100. Again, we trained two models with and without category information, originating from the clustering model learned on *late AMD categorization*.

IV. Results

We report quantitative and qualitative results that illustrate anomaly detection, visualize anomaly categorization outcome, provide descriptions of clusters according to experts and describe results of volume disease classification tasks using the identified categories as marker candidates.

A. Anomaly Detection Results

For the detection and segmentation of anomalies, the proposed method achieved a dice of 0.53 between annotated and

TABLE I
DICE, PRECISION, RECALL (=SENSITIVITY), SPECIFICITY, AND ACCURACY FOR ANOMALOUS REGIONS WITH MANUAL ANNOTATIONS, CALCULATED ON THE TEST SET. ADDITIONALLY, THE CHOSEN ν VALUE FOR THE ONE-CLASS SVM IS REPORTED.

Algorithm (ν)	Dice	Precision	Recall	Specificity	Accuracy
PCA_{256} *(0.4)*	0.47 (0.12)	0.36 (0.12)	0.74 (0.09)	0.46 (0.03)	0.55 (0.05)
$PCA_{0.95}$ *(0.2)*	0.51 (0.12)	0.40 (0.13)	0.74 (0.08)	0.56 (0.04)	0.62 (0.04)
$DCAE$ *(0.2)*	0.49 (0.13)	0.41 (0.13)	0.64 (0.14)	0.63 (0.07)	0.65 (0.05)
$\mathbf{DDAE_{ent}}$ ***(0.1)***	**0.53 (0.09)**	**0.47 (0.12)**	**0.63 (0.06)**	**0.71 (0.07)**	**0.69 (0.05)**

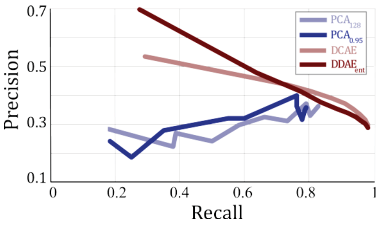


Fig. 4. precision-recall curve, calculated on the validation set.

predicted anomalous regions, a precision of 0.47, and a recall of 0.63, which means that 63% of all manually annotated anomalies were also identified as anomalous by our model (Table I). PCA_{256}, $PCA_{0.95}$ and $DCAE$ achieved a lower Dice (0.47, 0.51, and 0.49) compared to our method. Using a *paired Wilcoxon signed-rank test*, a significant difference could be shown for PCA_{256} (p=0.0004) and $DCAE$ (p=0.02), but not for $PCA_{0.95}$ (p=0.11).

To enable an objective qualitative evaluation, the volumes which are visualized in Fig. 3 were selected according to highest and lowest dice of $DDAE_{ent}$. An additional visual comparison of the segmentation results revealed that the shape of identified anomalous regions of the proposed method, $DDAE_{ent}$, reflected the manual annotations better than all comparison methods.

The validation performance for all examined ν values and all methods is reported in Fig. 4 and Fig. 6. At a recall level around 0.78, where the precision-recall curve (Fig. 4) seems to reveal comparable performance of the examined methods, $DDAE_{ent}$ achieves a precision of 0.42, outperforming all other approaches. At the same time, when comparing the curves, it can be clearly observed that both $DCAE$ and $DDAE_{ent}$ produced more stable results in comparison with the PCA methods. In particular, Fig. 6 shows that precision/recall decreased/increased continuously as increased for $DCAE$ and $DDAE_{ent}$, while both PCA methods exhibited an inconsistent behavior. In accordance with these quantitative outcome, Fig. 5 illustrates segmentation results for $DDAE_{ent}$ and $PCA_{0.95}$. Note that the embedding itself did not change with varying ν values. This inconsistency of both PCA methods makes an intuitive interpretation and adaption of ν difficult, though it may be important for specific tasks to control the precision-recall trade off.

B. Anomaly Categorization Results

Despite the fact that the anomaly detection performance left room for improvement in general, the detected anomaly candidates could be clustered into stable categories. The lowest DB-index was found for $C = 10$ on *late AMD categorization* and $C = 9$ on *early AMD categorization*, as indicated in Fig. 7(a). This was a plausible outcome, since OCT volumes with late AMD exhibit more obvious visual variation than early AMD volumes.

The cosine distance between cluster centroids is visualized in Fig. 7(b), where the columns were re-arranged for better interpretability. The nearest-neighbors of cluster centroids are illustrated in Fig. 7(c), both for *late AMD* and *early AMD clustering* results. As can be seen both in 7(b) and (c), all clusters of *early AMD clustering* could be linked to specific clusters in *late AMD clustering*. This was a plausible outcome, since all variation that is present in *early AMD*, is also present in *late AMD* cases. Exemplary category descriptions identified by experts are denoted in Fig. 7(d), where ”Upper boundary of photoreceptor layer with pathologic surrounding” (a4, b4), ”Photoreceptor layer with pathologic surrounding” (a5, b5), and ”vitreomacular interface with pathologic surrounding” (a9, b9) could be identified in both clusterings.

In contrast, *late AMD clustering* showed one additional cluster ”a10” which was identified as ”Exudative fluid” (e.g. intraretinal or subretinal fluid) segmentation by the clinical retina experts, and had no clear relation to a specific *early AMD* cluster. This claim of missing relation was supported by qualitative evaluation as well as by the calculated cosine distance between cluster centroids, which showed relatively low values (large distances) for ”a10” to all *early AMD* clusters, as illustrated in Fig. 7(b), bottom row. Clustering results are shown in Fig. 7 (e) on *late AMD test* B-scans, where it can be seen that cluster ”a10” showed substantial overlap with areas of fluid. Since fluid like intra- or sub-retinal fluid occurs only in late AMD, this was a reasonable outcome and indicated that also disease specific clusters had been learned.

C. Volume Level Disease Classification Results

We obtained an accuracy of 81.40% on the three-class classification task, using the volume of each cluster (corresponding to *late AMD clustering*) as features. The confusion matrix (Fig. 8(a)) shows that the classifier could successfully distinguish between late and early AMD cases. It is a more difficult task to separate early AMD and healthy volumes[2]. The feature importance, calculated during random forest training, is given in Fig. 8(b). It visualizes how each feature contributes to the prediction of a class in the form of the mean decrease of accuracy (MDA) for individual feature perturbations. We provide information about whether variables are positive or negative predictors by comparing their average value within class examples to the average value for out-of-class examples as the sign. Results identify ”a7” as the most important feature of the calculated random forest model. It is a strong negative predictor for healthy, while a strong positive predictor for late AMD. The comparison experiment without category information resulted in lower accuracy of 60.47% on the same classification task.

[2] Result examples can be found in the supplementary material.

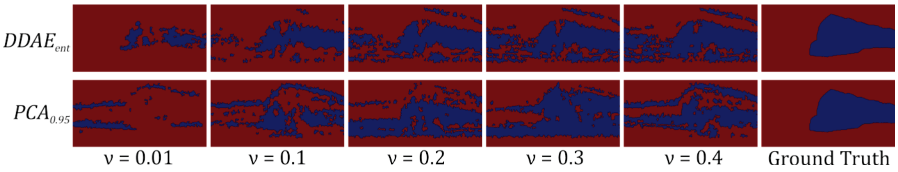


Fig. 5. Compared to the PCA methods, $DDAE_{ent}$ produced more stable results when varying ν. This finding was also supported by the segmentation results, illustrated for five consecutive ν values of $DDAE_{ent}$ and $PCA_{0.95}$ on an example B-Scan (anomaly regions are highlighted in blue).

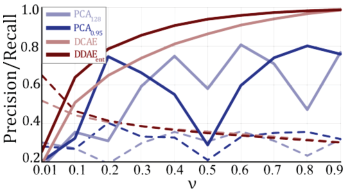


Fig. 6. Validation values of precision (dashed) and recall (solid) are plotted against ν.

On the *bioptigen test* set, we achieved an area under the ROC curve (AUC) of 0.944 (Figure 9) and 0.768 for the RF models using category information or not, respectively. The clear performance gaps on both classification tasks supports the claim that the learned anomaly categories encode clinically meaningful information. Furthermore, the result on *bioptigen test* indicates that the learned feature representation and the categories, respectively, reflect morphological properties of the retina, and not OCT vendor specific characteristics.

V. Discussion

We propose a method to detect and categorize anomalous regions in OCT volumes of the retina, and subsequently use these anomalies as marker candidates. The model is trained on healthy imaging data and detects anomalies in new volumes without constraints to a priori definitions. Categorization of anomalies revealed clusters of frequently occurring patterns, where a part of these categories could be mapped to clinically meaningful entities in the imaging data in a post hoc qualitative assessment of clusters by experts. Finally, results in disease classification tasks indicate that the identified marker candidates encode valuable discriminative information.

a) Three insights: From evaluation results we gain three primary insights. First, the proposed approach relying on a multi-scale deep denoising auto encoder architecture to represent image information shows comparable or superior performance to alternatives such as PCA or DCAE. At the same time, the embedding of $DDAE_{ent}$ allows to control the precision-recall trade off in an intuitive way, as opposed to PCA. This indicates that the representation is important for successful detection of subtle alterations in the imaging data and stable training of the one-class SVM.

Second, we can identify stable categories, that are replicable across data sets. Clustering reveals entities that are present in late- and early AMD, and a class of entities that is only present in late AMD. It demonstrates that purely data driven learning can reveal meaningful structure in the data, that corresponds to disease processes. Here, it reflects the emergence of exudative liquid that is characteristic for late AMD.

Third, the identified anomaly categories are valid marker candidates, that show predictive value, when used for volume level classification.

b) Relationship to prior work: While we achieved an AUC of 0.944 on the binary classification task, prior work reported an AUC of 0.984 [23] and 0.992 [21] on the Bioptigen dataset, where the latter used a different evaluation process (leave-one-out cross-validation on all cases). In [23] features are extracted at interest points which are located using manually defined constraints, while Farsiu *et al.* [21] used semi-automatically segmented retinal layers as features. In contrast to our study, both works use prior knowledge about the disease to create features specifically designed for this classification task. Additionally, our features are generated by a model which was trained on cases from a different OCT device (Spectralis vs. Bioptigen), which adds additional complexity to the task. Viewed in this light, our result indicates that the learned anomaly categories encode valuable discriminative information.

c) Identification of novel marker candidates: There is strong interest in the identification of valid biomarkers in AMD, since the already known biomarkers (e.g., retinal thickness, macular fluid) do not explain the entire spectrum of the disease and in particular the individual level of vision loss [6]. The proposed method contributes a path to identify novel marker *candidates*. It found categories that were known (e.g. photoreceptor layer with pathologic surrounding, cluster a5/b5), as well as potential new biomarker categories such as ”a7”, which could not clearly be linked to a particular known pathology by the clinical retina experts and at the same time showed a high feature importance regarding disease classification. The ultimate aim is to use unsupervised automated analysis to identify disease marker candidates in a first step, as done in this study, and to define a precise description of characteristics of those candidates in a second step, transforming them from candidates to effective markers applicable in clinical practice. The latter is subject of future work, for instance by correlating marker candidates with visual function. Results showing that the identified categories can classify disease are the strongest indication that unsupervised learning as proposed in this paper, can identify novel marker candidates and potentially contribute to understanding mechanisms governing disease course and treatment effect. If accuracy of the method can be improved further, in addition to marker identification, future work could also use anomaly detection to quickly visualize anomalies in OCT volumes, helping to efficiently evaluate large datasets, or in a screening setting.

d) Limitations: There are some limitations that have to be mentioned. First, the performance of the pixel-wise anomaly detection (dice=0.53) left room for improvement. While a recall of 0.63 indicates that manually annotated regions were still missed in this step, the relatively low precision of 0.47 may result from two sources: First, normal appearance dissimilar to the range represented in the training set, due to not having enough training data. The second possible source

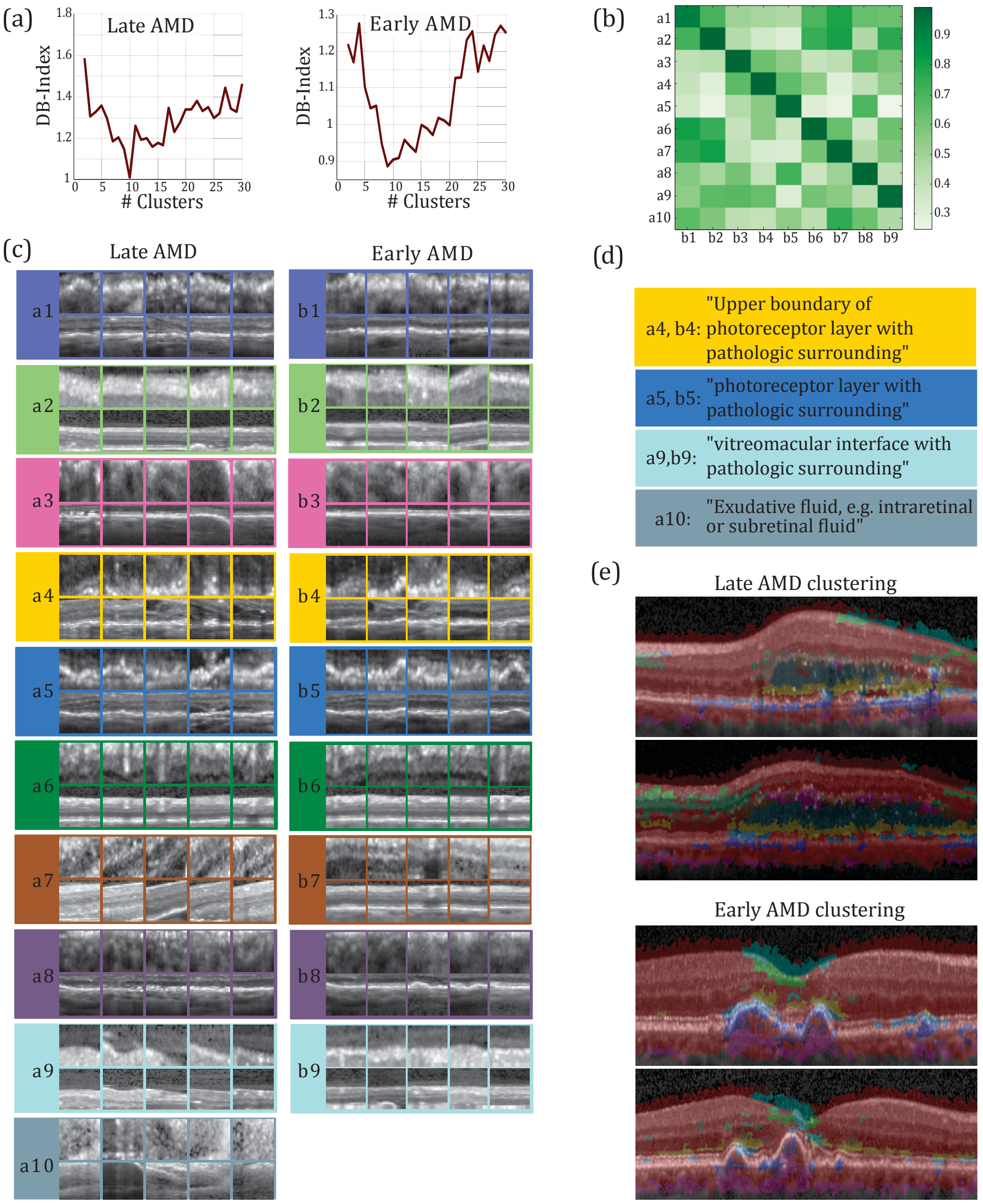

(a)
Late AMD
DB-Index
# Clusters
Early AMD
(b)
a1 a2 a3 a4 a5 a6 a7 a8 a9 a10
b1 b2 b3 b4 b5 b6 b7 b8 b9
0.9 0.8 0.7 0.6 0.5 0.4 0.3
(c)
Late AMD
Early AMD
a1 a2 a3 a4 a5 a6 a7 a8 a9 a10
b1 b2 b3 b4 b5 b6 b7 b8 b9
(d)
a4, b4: "Upper boundary of photoreceptor layer with pathologic surrounding"
a5, b5: "photoreceptor layer with pathologic surrounding"
a9,b9: "vitreomacular interface with pathologic surrounding"
a10: "Exudative fluid, e.g. intraretinal or subretinal fluid"
(e)
Late AMD clustering
Early AMD clustering

Fig. 7. Anomaly categorization: The calculated values of the DB-Index are plotted in (a). The cosine distance between cluster centroids is visualized in (b) and is bounded between 0 and 1, due to rescaling of features into positive domain. The nearest-neighbors of cluster centroids are illustrated in (c). The upper row shows the 32-by-32 patches, while the second row illustrates the 124-by-32 patches. Each cluster is indicated by a separate color. Some exemplary cluster descriptions that were identified by experts are denoted in (d). Clustering results of *late AMD* and *early AMD clustering* are shown in (e) on example B-scans, where identified anomalous regions were segmented into 10 and 9 categories, respectively. In accordance with former visualizations, normal regions are highlighted in red.

may be anomalies that have not yet been categorized, and are potential new candidates for markers. The interpretation of identified marker candidates remains challenging. They do not correspond to known categories, and thus no ground-truth exists for their direct evaluation. Instead we use expert description and classification experiments to verify and investigate their nature. Age information for individual patients was not available in this study. However, the datasets were composed of patients from multiple clinical studies, for which the average study age was available. The computed mean ages by weighting the mean ages of individual studies can be found in the supplementary material. In principle, a younger age of the healthy group could present a possible confounder in the biomarker identification and evaluation process. However, our data comes from clinical trials with a relatively high mean age (65.9 or higher). Additionally, signs of normal aging in OCT (i.e. mainly retinal pigment epithelium thinning [34]) are less pronounced than AMD related changes. Therefore, we expect that in this study, our method primarily picks up features associated with disease. A further limitation is that

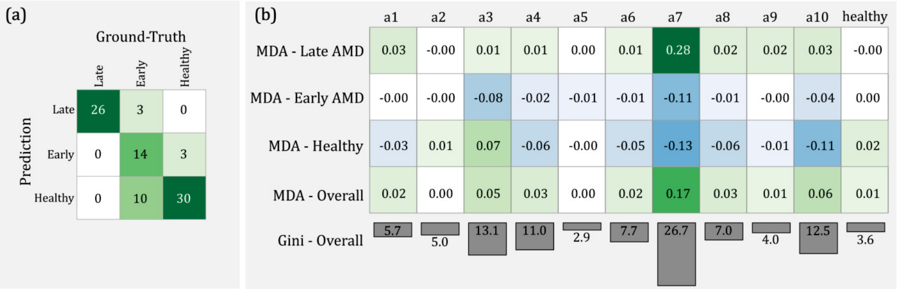


Fig. 8. Three-Class Classification: (a) Confusion matrix of the test set. (b) Identified anomaly categories as markers of disease. The first three rows show the class-specific MDA, where the sign encodes the feature-trend for that specific class (positive indicates high within class and low outside class values and vice versa). The fourth row contains the MDA over all classes, and the last row shows the mean decrease in Gini index.

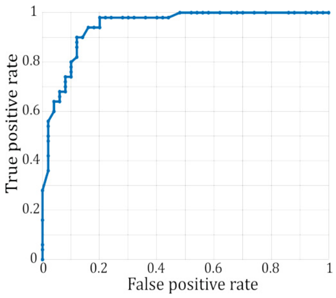


Fig. 9. ROC curve of the Bioptigen binary classification task.

contralateral OCTs of patients with RVO/AMD in the other eye were used as healthy training data. In order to minimize the influence of this potential bias, retina experts conducted careful selection of healthy OCTs within this data. There is a lack of scientific consensus regarding where normal aging of the retina stops and age-related disease starts. To address this limitation and to account for age-related changes that normally do not result in visual impairment, we have specifically included the mildest category of age-related changes which include small hard drusen ($< 63\mu m$) [8]. Another limitation is the restricted informative value of feature importance values in the case of low numbers of examples. A substantially higher number of decision trees is necessary to obtain stable feature scoring results compared to obtaining good classification accuracy. A further limitation is that the evaluation was conducted with AMD cases only, but, since the applicability of the proposed method is not limited to a specific anomaly, an extension to other diseases should be straightforward.

VI. Conclusion

We propose a method to segment anomalies in OCT volumes and categorize these findings into disease marker candidates. The detection of new anomalies, rather than the automation of expert annotation of known anomalies, is a critical shift in medical image analysis and particularly relevant in retinal imaging. In this context, we introduced a novel way to identify biomarker candidates, where results on both classification tasks indicate that valuable discriminative information is encoded in the newly identified categories. Future work is needed to transform these categories from candidates to actual markers applicable in clinical practice.

References

[1] R. Mayeux, “Biomarkers: potential uses and limitations,” *NeuroRx*, vol. 1, no. 2, pp. 182–188, 2004.

[2] 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.

[3] J. Fujimoto and E. Swanson, “The development, commercialization, and impact of optical coherence tomographyhistory of optical coherence tomography,” *Investigative Ophthalmology and Visual Science*, vol. 57, no. 9, 2016.

[4] 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,” *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[5] S. M. Waldstein, A.-M. Philip, R. Leitner, C. Simader, G. Langs, B. S. Gerendas, and U. Schmidt-Erfurth, “Correlation of 3-dimensionally quantified intraretinal and subretinal fluid with visual acuity in neovascular age-related macular degeneration,” *JAMA ophthalmology*, vol. 134, no. 2, pp. 182–190, 2016.

[6] U. Schmidt-Erfurth and S. M. Waldstein, “A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration,” *Progress in Retinal and Eye Research*, vol. 50, pp. 1–24, 2016.

[7] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014.

[8] F. L. Ferris, C. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, and S. R. Sadda, “Clinical classification of age-related macular degeneration,” *Ophthalmology*, vol. 120, no. 4, pp. 844–851, 2013.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[11] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs, “Predicting semantic descriptions from medical images with convolutional neural networks,” in *Information Processing in Medical Imaging*. Springer, 2015, pp. 437–448.

[12] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1422–1430.

[13] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2014, pp. 766–774.

[14] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun, “Stacked what-where auto-encoders,” *arXiv preprint arXiv:1506.02351*, 2016.

[15] Y. Zhou, D. Arpit, I. Nwogu, and V. Govindaraju, “Is joint training better for deep auto-encoders,” *arXiv preprint, arXiv: 1405.1380*, 2015.

[16] M. López, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, R. Chaves, P. Padilla, M. Gómez-Río, A. D. N. Initiative *et al.*, “Principal component analysis-based techniques and supervised classification schemes for the early detection of alzheimer’s disease,” *Neurocomputing*, vol. 74, no. 8, pp. 1260–1271, 2011.

[17] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, “High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning,” *Pattern Recognition*, 2016.

[18] D. Sidibé, S. Sankar, G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, G. S. Tan, D. Milea, E. Lamoureux, T. Y. Wong *et al.*, “An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images,” *Computer Methods and Programs in Biomedicine*, vol. 139, pp. 109–117, 2017.

[19] P. A. Dufour, H. Abdillahi, L. Ceklic, U. Wolf-Schnurrbusch, and J. Kowal, “Pathology hinting as the combination of automatic segmentation with a statistical shape model,” in *International Conference*

on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp. 599–606.

[20] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[21] 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.

[22] E. M. Lad, D. Mukherjee, S. S. Stinnett, S. W. Cousins, G. G. Potter, J. R. Burke, S. Farsiu, and H. E. Whitson, "Evaluation of inner retinal layers as biomarkers in mild cognitive impairment to moderate alzheimers disease," *PloS one*, vol. 13, no. 2, p. e0192646, 2018.

[23] F. G. Venhuizen, B. van Ginneken, B. Bloemen, M. J. van Grinsven, R. Philipsen, C. Hoyng, T. Theelen, and C. I. Sánchez, "Automated age-related macular degeneration classification in oct using unsupervised feature learning," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2015, pp. 94 141I–94 141I.

[24] P. Seeböck, S. Waldstein, S. Klimscha, B. S. Gerendas, R. Donner, T. Schlegl, U. Schmidt-Erfurth, and G. Langs, "Identifying and categorizing anomalies in retinal imaging data," *arXiv preprint arXiv:1612.00686*, 2016.

[25] M. K. Garvin, M. D. Abràmoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 9, pp. 1436–1447, 2009.

[26] M. Holzer and R. Donner, "Over-segmentation of 3d medical image volumes based on monogenic cues," in *Proceedings of the 19th CVWW*, 2014, pp. 35–42.

[27] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[28] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[29] K. Hornik, I. Feinerer, M. Kober, and C. Buchta, "Spherical k-means clustering," *Journal of Statistical Software*, vol. 50, no. 10, pp. 1–22, 2012.

[30] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107–145, 2001.

[31] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] F. Ko, P. J. Foster, N. G. Strouthidis, Y. Shweikh, Q. Yang, C. A. Reisman, Z. A. Muthy, U. Chakravarthy, A. J. Lotery, P. A. Keane *et al.*, "Associations with retinal pigment epithelium thickness measures in a large cohort: results from the uk biobank," *Ophthalmology*, vol. 124, no. 1, pp. 105–117, 2017.

CHAPTER 6

Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT

"Regardless of your faith,
you can never escape uncertainty."

– Shannon L. Alder

In this Chapter we exploit the potential of epistemic uncertainty, together with the concept of encoding anatomical knowledge into the model, for anomaly detection. Based on the assumption that epistemic uncertainty is high for structures that are not present in the training set, we train a segmentation model solely on healthy samples and exploit Bayesian deep learning (Section 3.2.4) to detect anomalous regions deviating from the normal training distribution in new images.

First, we use an existing automated method to segment normal anatomical structures in a healthy population, taking advantage of the fact that traditional segmentation methods are expected to perform accurately due to the well-defined properties of normal cases. Secondly, these *weak labels* of normal samples are used to train a U-Net (Section 3.2.2) on the segmentation task. On one hand, this means that the presented approach does not involve manual labels at any stage. On the other hand, this injects knowledge about normal anatomical variability into the model, implicitly incorporating information for detecting anomalies. At test time, epistemic uncertainty estimates (retrieved by using MC dropout sampling) are used to detect anomalous regions. Finally, a novel post-processing technique based on *majority-ray-casting* is applied in order to obtain smooth segmentations of the anomalies. We conduct an extensive evaluation of the proposed method on SD-OCT volumes from both healthy and diseased patients, including cases with late wet age-related macular degeneration (AMD), dry geographic atrophy (GA), diabetic macular edema (DME) and retinal vein occlusion (RVO).

The presented manuscript *"Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT"* has been submitted to the journal *"Transactions on Medical Imaging"* and is currently under review.

Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT

Philipp Seeböck*, José Ignacio Orlando, Thomas Schlegl, Sebastian M. Waldstein, Hrvoje Bogunović, Sophie Klimscha, Georg Langs*, and Ursula Schmidt-Erfurth

***Abstract*—Diagnosis and treatment guidance are aided by detecting relevant biomarkers in medical images. Although supervised deep learning can perform accurate segmentation of pathological areas, it is limited by requiring *a-priori* definitions of these regions, large-scale annotations, and a representative patient cohort in the training set. In contrast, anomaly detection is not limited to specific definitions of pathologies and allows for training on healthy samples without annotation. Anomalous regions can then serve as candidates for biomarker discovery. Knowledge about normal anatomical structure brings implicit information for detecting anomalies. We propose to exploit this property using bayesian deep learning, based on the assumption that epistemic uncertainties will correlate with anatomical deviations from a normal training set. A Bayesian U-Net is trained on a well-defined healthy environment using weak labels of healthy anatomy produced by existing methods. At test time, we capture epistemic uncertainty estimates of our model using Monte Carlo dropout. A novel postprocessing technique is then applied on these estimates to retrieve smooth segmentations of the anomalies. We experimentally validated this approach in retinal optical coherence tomography (OCT) images, using weak labels of retinal layers. Our method achieved a Dice index of 0.789 in an independent anomaly test set of age-related macular degeneration (AMD) cases. The resulting segmentations allowed very high accuracy for separating healthy and diseased cases with late wet AMD, dry geographic atrophy (GA), diabetic macular edema (DME) and retinal vein occlusion (RVO). Finally, we qualitatively observed that our approach can also detect other deviations in normal scans such as cut edge artifacts.**

***Index Terms*—weakly supervised learning, anomaly detection, biomarker discovery, optical coherence tomography, epistemic uncertainty.**

Manuscript received January,10,2019

The financial support by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development, by the Austrian Science Fund (FWF I2714-B31) is gratefully acknowledged. We thank the NVIDIA Corporation for GPU donation. J.I.O. is funded by WWTF AugUniWien/FA7464A0249 (Medical University of Vienna); VRG12-009 (University of Vienna).

P. Seeböck and G. Langs are with the Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Austria (email: philipp.seeboeck@meduniwien.ac.at, georg.langs@meduniwien.ac.at)

P. Seeböck, J. I. Orlando, T. Schlegl, S.M. Waldstein, H. Bogunovic, S. Klimscha, G. Langs and U. Schmidt-Erfurth are with the Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology and Optometry, Medical University Vienna, Austria. (email: philipp.seeboeck@meduniwien.ac.at)

* corresponding authors

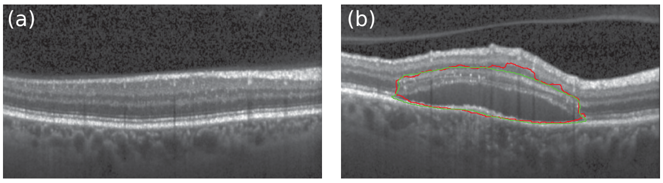


Fig. 1. Anomaly detection in retinal OCT. (a) Healthy retina. (b) Diseased subject (manual annotation of anomaly in green, prediction of anomaly by our model in red).

I. Introduction

Biomarker detection in medical imaging data plays a critical role in the context of disease diagnosis and treatment planning [1]. However, performing this task manually is extremely expensive and time consuming. Moreover, as it requires experts in the field to know every possible visual appearance of the regions of interest, results may suffer from intra- and inter-grader variability [2]. Automated methods can partially address these issues by exploiting the potential of deep learning [3]. Supervised learning approaches are trained to detect well-known, pre-defined biomarker categories such as lesions or pathological changes in organs and tissues [4]–[7]. In retinal OCT imaging, supervised methods have been extensively used [8], e.g. for segmentation of fluid [9], [10], drusen [11], hyperreflective material [12] or photoreceptor disruptions [13]. However, these methods require large-scale annotated data sets, which can be costly or even unfeasible to obtain in some clinical scenarios. Moreover, their outputs are limited to the pre-defined set of marker categories, and are unable to discover novel biomarkers different from those used for training [14].

Anomaly detection methods offer an interesting alternative to supervised learning in this domain, as they are not limited in their application to a specific disease or marker category. Instead, these approaches leverage the knowledge extracted from healthy data during training, omitting the need of a representative patient cohort with an appropriate amount and variations of pathologies [15], [16]. Capturing all possible disease related appearances or rare disease manifestations is costly or even unfeasible. In general, anomaly detection can be defined as a two-step process in which we first learn a model of normal appearance, and then we apply it to detect deviations from this normal data (anomalies) during test time [14], [15], [17]. Therefore, instead of searching in the entire image space, these segmented anomalies can be explored by clinicians to identify features that might result in novel biomarkers, allowing a more efficient discovery process. Furthermore, identifying anomalous areas can be also helpful to efficiently screen for diseased cases in large patient cohorts.

Bayesian deep learning has emerged as an active field of research that aims to develop efficient tools for quantifying uncertainties [18]–[20]. In general, uncertainties can be classified into two main categories: *aleatoric* and *epistemic*. *Aleatoric* uncertainty captures the vagueness inherent in the input, while *epistemic* uncertainty refers to the model incertitude and can be reduced by incorporating additional data into the training process [20]. Both aleatoric and epistemic uncertainty have previously been used for semantic segmentation [19], [20]. In this context it has been shown that aleatoric uncertainty does not increase for examples different from those in the training set, while epistemic uncertainty does [20]. Hence, the latter is more suitable for detecting changes (or anomalies) from the normal samples. Furthermore, disease classification methods based on deep learning were observed to be benefited by the usage of epistemic uncertainties [21].

In this paper, we introduce a novel approach for anomaly detection, exploiting segmentation models of normal anatomy and their epistemic uncertainty while segmenting new images. Our method is based on the assumption that these uncertainties will correlate with deviations from a normal appearance. We learn the regularities in the anatomy of healthy data, using weak labels. In this work we use the term ”weak supervision” to indicate that we trained our model using labels automatically generated by a surrogate segmentation method instead of a human reader. We exploit this characteristic as traditional algorithms–even if they are not based on machine learning–are expected to perform accurately due to the well-defined properties of normal cases. Therefore, our approach does not involve manual labels at any stage. This setting allows to produce more training data and thereby to harvest more appearance variability.

We experimentally evaluate our approach in the context of anomaly detection in retinal optical coherence tomography (OCT) scans (Fig. 1, Section I-A). We train a Bayesian U-Net [22], [23] on a set of healthy images using weak labels of the retinal layers, provided by a standard graph-based method for layer segmentation [24]. At test time, we capture the epistemic uncertainty estimates from our network by means of Monte Carlo (MC) dropout [18], [19]. This output is postprocessed using a novel *majority-ray-casting* technique in order to retrieve compact, blob-shaped smooth segmentations of the anomalies. On a separate test set of patients with age-related macular degeneration (AMD), our method achieves a Dice index of 0.789, outperforming previously published work by a large margin. Furthermore, the performance of the proposed method is evaluated in a volume-level classification experiment, using only the amount of anomalous area as (discriminative) feature. By individually comparing healthy cases vs. diabetic macular edema (DME), retinal vein occlusion (RVO), dry geographic atrophy (GA) and late wet AMD, we observe that even this simple predictor allows to achieve almost perfect separation.

A. Retinal OCT imaging

OCT is a non-invasive volumetric imaging technique that provides high resolution images of the retina and is currently one of the most important diagnostic modalities in ophthalmology [25]. A 3D OCT volume is composed of several 2D cross-sectional slices–or B-scans–, which are analyzed by physicians to determine treatments, diagnosis and other clinical decisions [25]. Age-related macular degeneration (AMD) is one of the leading causes of blindness in the world [26]. Detectable AMD-related changes in OCTs are, among others, drusen, intra- and subretinal fluid, pigment epithelial detachment (PED) and photoreceptor loss [8]. Besides neovascular AMD, which is defined by the occurrence of fluid, geographic atrophy (GA) is the second form of late AMD, characterized by the death of retinal pigment epithelium (RPE) cells, photoreceptors and/or choriocapillaris. Other retinal diseases such as retinal vein occlusion (RVO) [27] and diabetic macular edema (DME) [28] are characterized by the occurrence of intraretinal/subretinal fluid. Presence or changes in some of these features have been shown to correlate with visual function or disease progression [29]. Predictive capability however remains to be limited and underlying pathogenetic mechanisms are not yet fully understood [30], meaning that there might be other unknown structures or patterns that are still needed to be discovered.

We propose to apply our uncertainty based approach to automatically segment anomalies in retinal OCT scans. In this domain, *normal* is defined as the absence of pathological changes beyond age-related alterations. According to the Beckman Initiative Classification [31], we allowed drusen below 63 μm in size as only visible alteration, as they normally do not result in visual impairment. A set of healthy retinas and corresponding weak labels obtained using [24] are used to train a Bayesian deep learning model for segmenting the retinal layers. Pixel-wise epistemic uncertainty estimates are applied at test time to identify anomalous regions in new given samples. While pathologies such as subretinal fluid are known to alter the appearance of the retina, some other are strictly related with the layers (e.g. the disorganization of the retinal inner layers, or DRIL) [23]. Therefore, using retinal layer information is an appropriate way of incorporating anatomical knowledge into the model. At the same time, no labels of the target class (i.e. anomalies) are needed for training.

B. Related Work

Biomarker discovery and analysis have benefited by the incorporation of deep learning [32]. Non data-driven approaches require hand-crafting techniques to capture a specific biomarker, and then assess its statistical power, e.g. by means of linear discriminant analysis [33]. Alternatively, supervised deep learning avoids biases due to manual design of features by learning them from data. These techniques have been extensively used to identify pre-defined pathological markers such as disease lesions [4], [7], [8], [34]. Their main drawback is that they require a training set with manual annotations of the region of interest. Thus, the markers have to be known in advance–restricting the possibility of using these models

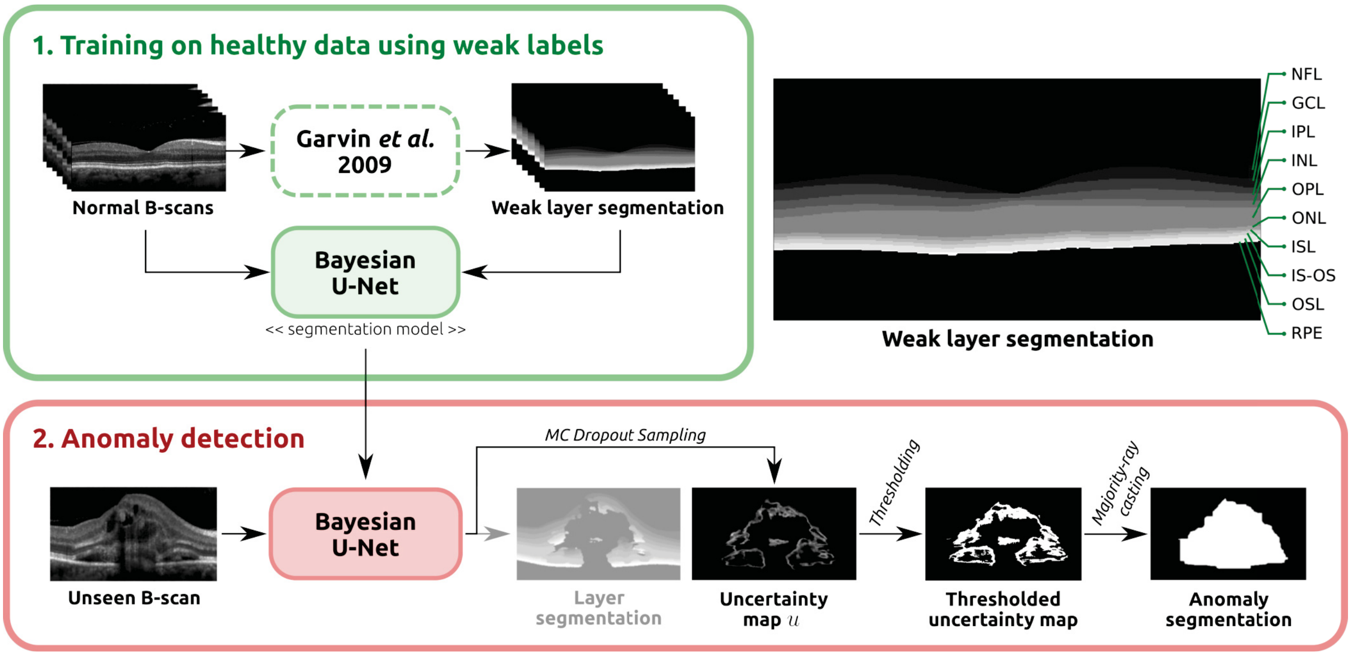


Fig. 2. Overview of the proposed method. A Bayesian U-Net is trained on *normal* B-scans, using weak labels of the retinal layers generated using the Garvin *et al.* [24] segmentation method. The retinal layers are indicated on the right hand side. Given an unseen B-scan, MC dropout sampling is used to retrieve epistemic uncertainty maps, which are subsequently post-processed using *majority-ray-casting* to obtain the final anomaly segmentations.

for biomarker discovery–and the training set has to include enough representative examples of the marker appearances. Alternatively, some authors propose to predict pre-defined clinical parameters using image-based regression techniques [32], [35]. These methods assume that the networks will learn to capture features in the images that are correlated with relevant target values. Appropriate visualization techniques are needed to understand the properties of the model and to identify the features taken into account for prediction [32], [35]. The regression target needs to be pre-defined, and it can be either a functional parameter [32] or a diagnosis [35]. Furthermore, a representative sample of diseased subjects has to be included in the training set if the target parameters are related to a specific condition. Moreover, due to the complexity of the prediction task, a larger number of training samples is required compared to supervised segmentation approaches.

Anomaly detection, on the contrary, identifies pathological areas that are implicitly defined by healthy data: normal appearance is first learned from this data, and anomalies are obtained in new data by detecting the difference to this representation. This overcomes the need of a sufficiently representative cohort of diseased patients, to select features with stable predictive value for a given target. Instead, first anomalies are detected based on a model trained on large-scale healthy data, and highlighted in the images as blob-shaped segmentations. In a second step, these candidates–typically only a fraction of the overall data–can be mined more efficiently for discovering new biomarkers and/or predictors. These techniques can be applied as a first step in discovering novel risk factors of diseases, extending the vocabulary of known biomarkers, and therefore our knowledge about the underlying pathogenesis of diseases [14], [17], [36].

Multiple techniques have been proposed in the past for automated anomaly detection in OCT images [14], [17], [37], [38]. Shape models were used in [37] to perform drusen detection. In [38], the appearance of normal OCT B-scans was modeled with a Gaussian Mixture Model (GMM), recognizing anomalous B-Scans as outliers. Entire OCT volumes were classified as normal or anomalous, based on the number of outliers. Deep unsupervised anomaly detection has been recently presented in [14], [17], both relying on a representation learned at patch-level. Schlegl *et al.* [17] used a Generative Adversarial Network (GAN) to learn a manifold of normal anatomical variability, and anomalies were detected as deviations from it. A multi-scale autoencoder approach combined with a one-class support vector machine (SVM) was presented in [14] to segment anomalies and to identify disease clusters subsequently. None of these anomaly detection approaches incorporate the use of uncertainty.

To the best of our knowledge, uncertainties were not used for anomaly detection before. In particular, Nair *et al.* [34] used Bayesian supervised learning to segment multiple sclerosis lesions in MRI. Sedai *et al.* [39] applied a similar method for layer segmentation in healthy OCT scans. In both works, aleatoric uncertainty was used for training. In [34], epistemic uncertainty was applied to refine the segmentations, while in [39] the epistemic uncertainty was provided as qualitative feedback to users. Monte Carlo sampling with dropout was used in [36] to average multiple outputs from an autoencoder trained in healthy data. Anomalies were detected as differences between the input and the reconstructed output. In this paper we aim for a different task compared to these previous approaches: we use the epistemic uncertainty of a model trained on healthy subjects to discover anomalies in new data.

C. Contributions

We propose a novel approach for anomaly detection based on epistemic uncertainty estimates from a Bayesian U-Net, trained for segmenting the anatomy of healthy subjects. To the best of our knowledge, this is the first method to pose the segmentation of anomalies in this way. In addition, our model is trained on weak labels instead of manual annotations, which allows to increase the training data without major efforts. We evaluate our model in the context of anomaly detection in retinal OCT scans. We introduce a heuristical postprocessing technique, namely majority-ray-casting, to ensure compact-shape consistency in the final anomaly segmentations. Our approach is able to obtain a clear performance improvement compared to previous state-of-the-art in anomaly detection in OCTs [14]. This manifests in a Dice of 0.789 on the anomaly test set regarding the pixel-wise segmentation task, while achieving almost perfect volume-level separation of healthy and diseased volumes with late wet AMD, dry GA, DME and RVO, solely based on the area of detected anomalies. Finally, we also qualitatively observed high uncertainty estimates in regions with other deviations such as imaging artifacts in normal subjects.

II. Methods

An overview of the proposed approach is illustrated in Fig. 2. First, we train a Bayesian U-Net model on *normal* cases to segment retinal layers, using weak labels automatically generated with a graph-based segmentation approach. Secondly, this model is applied together with Monte Carlo dropout [18], [19] to retrieve pixel-level epistemic uncertainty estimates. Finally, we introduce a simple post-processing step, *majority-ray-casting*, to transform the uncertainty maps into compact segmentations of anomalies. This technique closes the gap between the shape of layers and anomalies based on the assumption that anomalies in OCT are compact and not layered.

Section II-A describes the general idea of training a segmentation model from a healthy population using weak labels. Section II-B focuses on the application of the epistemic uncertainty estimates of this model for anomaly detection. The domain-specific pipeline for applying the anomaly detection approach in retinal OCT scans is presented in Section II-C

A. Training on Healthy Population

Let $X \in \mathbb{R}^{a \times b}$ be a set of *normal* images with $a \times b$ pixels size, and $Y \in \mathcal{Y}^{a \times b}$ the set of corresponding weak, target label maps, with $\mathcal{Y} = \{1, ..., K\}$ the set of all possible classes. A segmentation model aims at finding the function $f_W : X \rightarrow Y$ by optimizing its set of weights W. In this study, we model f_W using a multiclass U-Net [22]. This widely used segmentation architecture is composed of an encoding and a decoding part with skip-connections: the encoder contracts the resolution of the input image and captures the context and relevant features on it, while the decoder performs up-sampling operations to enable precise localization of the target class and restores the input resolution. The skip-connections, on the other hand, allow to better reconstruct the final segmentation by transferring feature maps from one encoding block to its counterpart in the decoder. Our instance of the U-Net (Fig. 3) comprises five levels of depth, with 64, 128, 256, 512 and 1024 output channels each. Dropout is applied after each convolutional block, which consists of two 3×3 convolutions, each followed by batch-normalization [40] and a rectified linear unit (ReLU). 2×2 max-pooling and nearest-neighbor interpolation are used for downsampling and upsampling, respectively. The network is trained with the cross entropy loss objective function.

B. Exploiting Epistemic Uncertainty for Anomaly Detection

Epistemic uncertainty was observed to increase when estimated on image samples whose appearance differ significantly from those on the training data [20]. We propose to exploit this characteristic to identify and segment anomalies in unseen scans.

Formally, Bayesian deep learning aims to find the posterior distribution over the weights of the network $p(W|X, Y)$, in order to derive epistemic uncertainty. In general, retrieving the actual true underlying distribution is computationally intractable, so it needs to be approximated. Gal *et al.* [18] proposed to approximate the posterior with the variational distribution $q(W)$, i.e. by using dropout also at test time to retrieve MC samples. This is theoretically equivalent to modelling q as a Bernoulli distribution with probability p equal to the dropout rate. It has been shown in [18] that the Kullback-Leibler divergence between the approximate and posterior distribution:

$$KL(q(W)||p(W|X, Y)) \tag{1}$$

is minimized by optimizing the cross-entropy loss during training. Hence, training the network with gradient descent and dropout not only prevents over-fitting but also encourages the network to learn a weight distribution that properly explains the training data.

At test time, given an *unseen* image x (e.g. a B-scan), the pixel-wise epistemic uncertainty is estimated as follows. First, n predictions $\mathbf{y}^{(i)}$, $i \in 1, \dots, n$ are retrieved by applying the model $f_{W \sim q(W)}$ on x. The pixel-wise variance σ^2 is then computed for each class $k \in \mathcal{Y}$ by:

$$\sigma_k^2(\mathbf{p}) = \frac{1}{n} \sum_i^n \left(y_k^{(i)}(\mathbf{p}) - \mu_k(\mathbf{p}) \right)^2 \tag{2}$$

where $\mathbf{p}$ is a pixel coordinate and μ_k is the average of the n predictions for the k-th class. The final uncertainty map u is obtained by averaging all σ_k^2 estimates over the K class-specific variances in a pixel-wise manner:

$$u(\mathbf{p}) = \frac{1}{K} \sum_k^K \sigma_k^2(\mathbf{p}). \tag{3}$$

C. Application of anomaly detection in retinal OCT scans

We apply the uncertainty-based anomaly detection approach to retinal OCT scans. The training set consists of pairs (X, Y) composed of a healthy OCT B-scan X and its associated weak labelling map Y of the retinal layers. Y is pre-computed

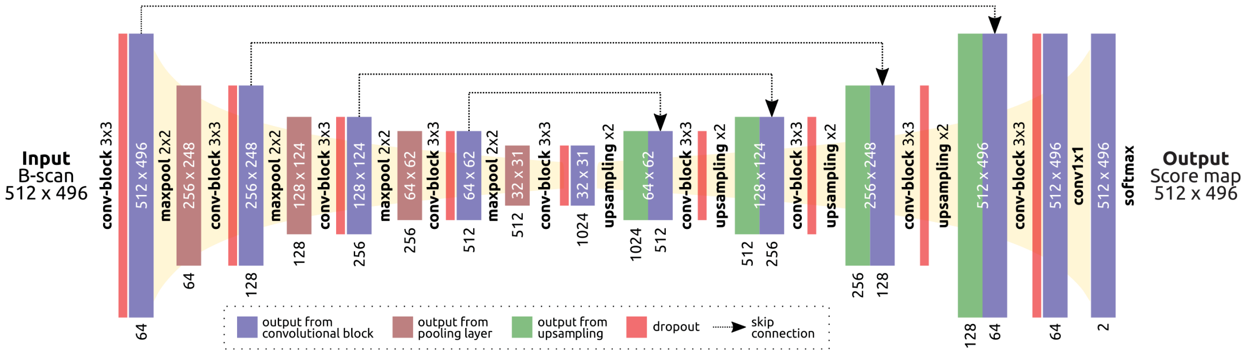


Fig. 3. Overview of the network architecture. Each convolutional block has the following structure: 3-by-3 convolution + batch-normalization + ReLU + 3-by-3 convolution + batch-normalization + ReLU. All convolutional layers use a stride of 1 and zero padding. A combination of nearest neighbor upsampling and a convolutional layer is applied instead of transposed convolutions.

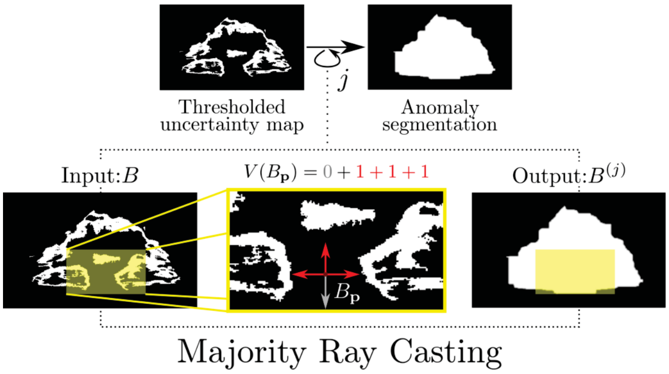


Fig. 4. Majority-ray-casting postprocessing technique. Red arrows indicate ray hits with a binary (white, =1) region, while the gray arrow indicates a non-hit.

using the graph-based surface segmentation algorithm described in [24]. Such a method has proven to be effective in normal subjects and is widely applied in ophthalmological studies [41], [42]. The set $\mathcal{Y}$ of labels comprises $K = 11$ classes corresponding to background and 10 retinal layers (Figure 2): nerve fiber layer (NFL); ganglion cell layer (GCL); inner plexiform layer (IPL); inner nuclear layer (INL); outer plexiform layer (OPL); outer nuclear layer (ONL); inner segment layer (ISL); inner segment - outer segment (IS-OS) junction; outer segment layer (OSL) and the retinal pigmented epithelium (RPE).

We use these weak labels to train the Bayesian multiclass U-Net described in Section II-A. The neural network provides both a segmentation map and an uncertainty estimate. We only use the latter at test time, as our purpose is not to accurately identify the retinal layers but to segment retinal abnormalities.

A first estimate of the anomalous areas is obtained by thresholding u with a threshold t. To eliminate spurious predictions, every connected component with an area smaller than s pixels is removed, resulting in a binary map B. The most straightforward way to highlight anomalies in an input B-scan is by providing compact, blob-shaped smooth segmentations surrounding the abnormal areas. As can be seen Fig. 4, B is not smooth enough to fit that shape.

We introduce a simple but effective technique, *majority-ray-casting*, that iteratively postprocesses the binary map B and results in a more shape consistent anomaly segmentation. This approach assumes that the retina is approximately horizontally orientated in the B-scan, which is usually the case. A schematic representation of the method is provided in Fig. 4. On an iteration j, in a first step four rays are sent to each of the cardinal coordinates (left, right, top and bottom) from every pixel $\mathbf{p}$ that satisfies $B_{\mathbf{p}} = 0$. In other words, each black pixel in Fig. 4 is used once as reference point to cast the four rays. Each ray that "hits" a pixel with value 1 before reaching the border of B increases a pixel-wise ray-casting vote $V(B_{\mathbf{p}})$ by 1. Hence, the maximum voting value of $V(B_{\mathbf{p}})$ for each pixel $\mathbf{p}$ can be 4. In a second step, all pixels with votes greater than or equal to a hyper-parameter $v^{(j)}$ are then set to 1, resulting in a new binary map $B^{(j)}$. Formally, this can be written as:

$$B_{\mathbf{p}}^{(j)} = \begin{cases} 1 & \text{if } B_{\mathbf{p}} = 1 \\ 1 & \text{if } V(B_{\mathbf{p}}) \geq v^{(j)} \\ 0 & \text{if } V(B_{\mathbf{p}}) < v^{(j)}. \end{cases} \tag{4}$$

Notice that this process can be iteratively repeated using $B^{(j)}$ as an input to the next iteration, and a different value of $v^{(j)}$ can be used at each iteration. Finally, morphological closing and opening operations with a radius of m_c and m_o, respectively, were applied to remove artifacts.

III. Experimental Setup

We empirically evaluated our method in our application scenario. In particular, we studied: (1) if our method can accurately identify anomalous regions in retinal OCT data, (2) the contribution of each of the individual components of our proposed approach in the final results, (3) the lesion-wise detection performance of the method, and (4) the volume-wise classification accuracy of the algorithm, based on the average number of anomalous pixels per B-scan for each volume..

a) Data: We used six data sets of macula centered Spectralis (Heidelberg Engineering, GER) OCT scans, with 512 × 496 × 49 voxels per volume, covering approximately 6mm × 2mm × 6mm of the retina. The first two datasets *normal* and *normal evaluation* comprise 226 and 33 healthy volumes, respectively, which were selected from 482 / 209 contralateral eye scans of patients with Retinal Vein Occlusion (RVO) / AMD in the other eye. According to the definition of *healthy* provided in Section I, volumes with pathological changes beyond age-related alterations were excluded. The *normal* data set was randomly split on a patient basis into 198 training and 28 validation cases to train our segmentation model. The *normal evaluation* set, on the other hand, was only used for evaluation purposes. The third dataset *late wet AMD* comprised 31 OCT volumes (5 validation, 26 test) with active neovascular AMD. A retina specialist manually annotated all the areas containing pathologic features, resulting in pixel-wise annotations of anomalous regions. All these datasets have been already used for training and evalution in [14], using the same configuration. This allows a direct comparison with such an approach.

Four volume-wise disease classification experiments were performed by comparing the anomalous areas in healthy subjects vs diseased. The *late wet AMD test set*, 30 DME, 25 RVO and 34 dry GA volumes were used separately for this purpose.

b) Training details: Intensity values of each individual B-scan were rescaled between 0 and 1 before being fed to the network. We used Kaiming initialization [43], Adam optimization [44], and a learning rate of 1×10^{-4} (multiplied by 0.2 every 5 epochs). During training, random data augmentations were applied, including horizontal flipping, rotation up to $10°$, horizontal / vertical translations up to 5% / 20% of the B-scan size and scaling up to 2%. The network was trained for 25 epochs on the *normal training set*, and the model with the best average Dice for layer segmentation on the *normal validation set* was selected for evaluation. We trained the model with different dropout rates $p = \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and the model with the highest Dice for anomaly segmentation on the *late wet AMD validation set* was selected for performance evaluation on the *late wet AMD test set*.

c) Anomaly detection details: At inference time, 50 MC samples with dropout were retrieved per B-scan. For post-processing, we used $s = 10$, $m_c = 4$, and $m_o = 2$, where these parameters were selected empirically by qualitatively analyzing the results in a few B-scans from the *late wet AMD validation set*. Two iterations of the *majority-ray-casting* algorithm were performed, using $v^{(1)} = 3$ and $v^{(2)} = 4$, and 20 different thresholds $t = \{0.01, 0.02, ...0.19, 0.20\}$ were evaluated on the *lateAMD validation set*. The best threshold according to the average validation Dice was selected for performance evaluation on the *lateAMD test set*. This calibration ensured to retrieve compact annotations consistent with the desired blob-shape appearance.

A. Segmentation accuracy

The segmentation accuracy was evaluated using precision, recall and Dice, which are standard metrics for binary segmentation tasks. Notice that performing a ROC curve based evaluation is unfeasible in our case as our method does not produce pixel-level likelihood predictions of anomalies, but binary labels.

To assess the contribution of each individual component of our proposed approach in the final results, we performed a series of ablation experiments. It is worth mentioning that the test set was not used for designing the method: all our design decisions were based on the validation set performance. These ablation studies are performed on the test set only to illustrate how changing our model can affect the results. For the sake of brevity, from now on we will refer to the full method described in Section II as *WeakAnD* (from ***Weak An****omaly* ***D****etection*).

- *Binary layer-segmentation*: While the proposed *WeakAnD* is trained with 11 layer classes, we trained a second network, namely *WeakAnD(binary)*, for the binary segmentation task ”retina/background”. This experiment allows to assess the influence of annotation details in the anomaly detection performance.
- *Remove majority-ray-casting*: To show the necessity of the majority-ray-casting approach, we compared against a simple post-processing only thresholding the uncertainty maps u (*WeakAnD (thresholding)*). We also replaced the majority-ray-casting step with a straightforward convex hull step (*WeakAnD (convex-hull)*).
- *Remove morphological operations*: The final morphological closing and opening operations were removed in this ablation experiment (*WeakAnD (w/o closing/opening)*).
- *Layer flattening*: As an additional pre-processing step for the *lateAMD* dataset, the retina was flattened using the bottom layer (Bruch’s Membrane - BM), projecting it onto a horizontal plane, following the pre-processing approach in [14]. Our hypothesis is that flattening the retina helps to meet the assumption of majority-ray-casting, i.e. horizontal orientation of the retina.

B. Lesion-wise Detection

We are interested in evaluating the detection performance of the proposed approach on a lesion-wise basis. To this end, we define each connected anomaly within a B-scan as a single lesion (e.g., Fig. 7(b) presents two lesions). A Dice index is computed for each individual lesion to quantify its overlap with its corresponding manual annotation. A thresholding according to a reference value d is then performed, where the amount of true positives is counted as the number of lesions with a Dice index higher than d. These values are used to compute *lesion-detection Recall (LD-Re$_d$)* and *lesion-detection Precision (LD-Pr$_d$)*:

$$\text{LD-Re}_d = \frac{\text{TP}_d}{\text{TP}_d + \text{FN}_d} \tag{5}$$

$$\text{LD-Pr}_d = \frac{\text{TP}_d}{\text{TP}_d + \text{FP}_d} \tag{6}$$

where TP_d, FN_d and FP_d are the number of true positive, false negative and false positive lesions for a given d. By computing these metrics for each possible $d \in [0, 1]$, we can then plot both *LD-Re* and *LD-Pr* curves. These plots allow to assess the stability of the Dice values with respect to the lesion detection performance. Notice that this cannot be used to select an operating point as it is defined over all possible dice values and not on lesion probabilities.

C. Volume-wise Disease Detection

We conducted four additional experiments to evaluate if the proposed method can be used to discriminate diseased

TABLE I
QUANTITATIVE RESULTS ON THE LATE WET AMD VALIDATION SET WITH VARYING DROPOUT PARAMETERS.

Dropout	Dice
0.1	0.783 (0.05)
0.2	0.778 (0.02)
0.3	0.796 (0.03)
0.4	0.798 (0.03)
0.5	0.783 (0.03)

TABLE II
QUANTITATIVE RESULTS OF ANOMALY DETECTION ON THE LATE WET AMD TEST SET.

Method	Precision	Recall	Dice
DDAE$_{\text{ent}}$ [14]	0.47 (0.12)	0.63 (0.06)	0.53 (0.09)
Entropy of Soft Predictions	0.600 (0.08)	0.622 (0.09)	0.606 (0.07)
WeakAnD	**0.739 (0.06)**	**0.808 (0.07)**	**0.768 (0.03)**
WeakAnD (with layer-flattening)	**0.748 (0.06)**	**0.844 (0.07)**	**0.789 (0.03)**

versus healthy patients and to further assess the behavior of our approach on healthy cases. Without additional training, the average anomalous area per B-scan for each volume was directly used as a discriminative feature to separate between healthy and diseased cases. The following setups were used: *normal evaluation* vs *late wet AMD test set*, *normal evaluation* vs GA, *normal evaluation* vs RVO and *normal evaluation* vs DME.

IV. RESULTS

Quantitative results for anomaly detection are provided in Table II. Two baselines are included: the state-of-the-art method described in [14] and an additional approach based on replacing our epistemic uncertainty estimates by the entropy of the soft predictions of the layers. It can be seen that the proposed approach outperformed the two baselines by a large margin. When layer-flattening is applied to pre-process the OCT volumes as in [14], an improvement in performance is also observed, with a statistical significant increment in the Dice values from 0.768 to 0.789 (paired Wilcoxon signed-rank test, $p = 0.00007$). The final *WeakAnD* model used a threshold of $t = 0.10$ and a dropout rate of $p = 0.4$. However, we experimentally observed that the performance on the validation set was not too sensitive to the dropout parameter (Table I.

Qualitative anomaly segmentation results obtained in the *late AMD test set* are shown in Fig. 5. The central B-scans, corresponding to the volumes in which our method performed best/worst in terms of Dice, are shown in the top/bottom two rows. An additional example of a non central B-scan is depicted in Fig. 1. Further qualitative results in DME, RVO and GA cases are depicted in Fig. 11 and in the supplementary material.

A scatter plot comparing the total area (in pixels) of anomalies (as manually annotated by the expert) and the level of uncertainty of the segmentation model is depicted in Fig. 6. Each point corresponds to an individual OCT volume in the *late wet AMD test set*. The linear regression line for the corresponding values is also included in the plot. The correlation between variables, as measured using the Pearson correlation coefficient, is $\rho = 0.91$.

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION STUDIES, AS EVALUATED ON THE LATE WET AMD TEST SET.

Method	Precision	Recall	Dice
WeakAnD (thresholding)	0.614 (0.05)	0.504 (0.06)	0.550 (0.04)
WeakAnD (binary)	0.716 (0.07)	0.620 (0.12)	0.655 (0.07)
WeakAnD (convex-hull)	0.708 (0.07)	0.836 (0.08)	0.761 (0.04)
WeakAnD (w/o closing/opening)	0.727 (0.06)	0.815 (0.07)	0.765 (0.03)
WeakAnD	**0.739 (0.06)**	**0.808 (0.07)**	**0.768 (0.03)**

a) Segmentation Accuracy: Table III provides quantitative results of the conducted ablation studies, while qualitative results are shown in Fig. 7. It can be observed that all the ablations resulted in a performance loss, with different quantitative and qualitative effects. In particular, the importance of using a fine-grained layer segmentation is highlighted by the drop in the observed evaluation metrics when using a binary segmentation.

b) Lesion-wise Detection: Lesion-wise precision and recall curves are shown in Fig. 8. The corresponding curves for the baseline methods are also included for comparison purposes.

c) Volume-wise Disease Detection: Fig. 9 depict histograms for the volume-wise disease detection experiment, both for the two baselines ([14] (a) and entropy (b)) and our method (c). Red bars correspond to patients from the *late wet AMD* data set, while green bars are associated to patients in the *normal evaluation* set. The horizontal axis represents the average number of anomalous pixels per B-scan for each volume, while the vertical axis indicates the number of patients with a similar anomalous area. Fig. 9 (b) and (c) shows no overlap between the healthy and the abnormal sets, while Fig. 9 (a) does. Qualitative examples of the anomalies detected in healthy cases from the *normal evaluation* set are depicted in Fig. 12. Both images correspond to the cases with the largest anomalous area. The detected anomalies in these cases correspond to imaging artifacts (Fig. 12, top) or small deviations from normal retinas such small drusen deposits (Fig. 12, bottom). A small false positive is observed at the center of the fovea.

Fig. 10 presents scatter plots showing the average number of anomalous pixels per B-scan for each diseased/healthy volume in our volume-wise classification experiments. As in Fig. 9, it can be seen that this feature is an almost perfect predictor for this application. Qualitative results of the central B-scan of DME, RVO and GA cases, respectively, are presented in Fig. 11. The anomalous region detected in Fig. 11 (a) covers parts of the retina with intraretinal cystoid fluid. The segmentation in Fig. 11 (b) shows a similar behaviour, although it also includes areas of intraretinal hyperreflective foci. Finally, Fig 11 (c) illustrates that our method is also capable of selectively detecting areas of RPE atrophy and neurosensory thinning in eyes with GA.

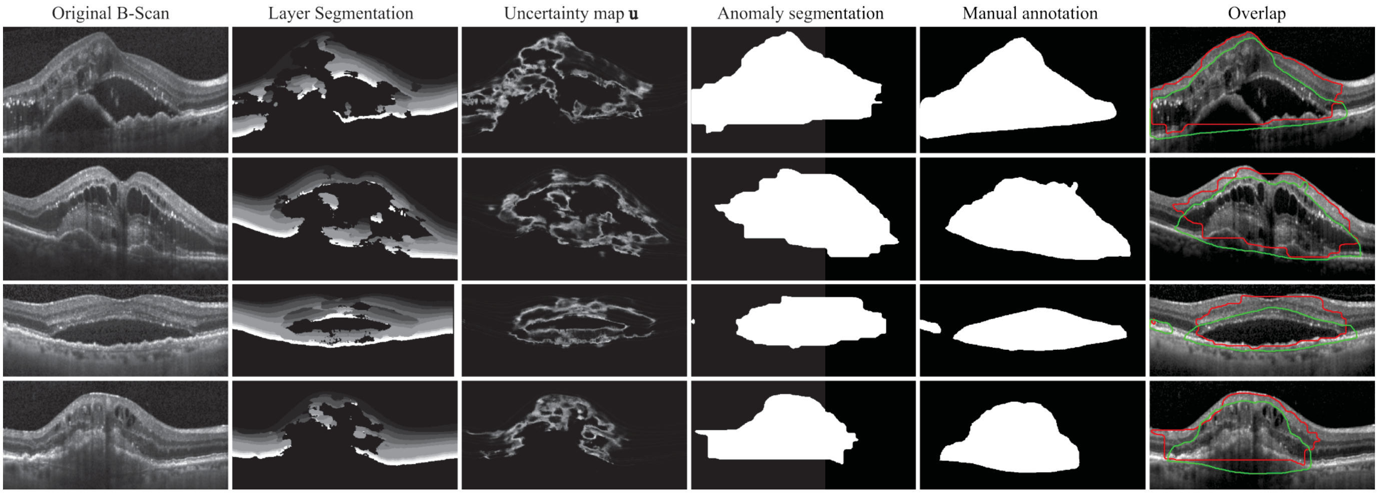


Fig. 5. Qualitative results of the proposed method, on the late wet AMD test set. Central B-scans of volumes in which the proposed method performed best/worst in terms of the Dice index are shown. The corresponding Dice values are 0.82, 0.81, 0.72 and 0.72, from top to bottom. The last column indicates the overlap between the manual annotations of anomaly in green and the prediction of anomaly by our model in red.

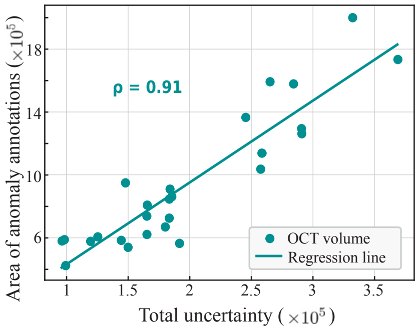


Fig. 6. Correlation between the total amount of uncertainty and the area of anomaly annotations. Each point in the plot corresponds to one OCT volume in the *late wet AMD test set*. The least-squared-fit line as well as the Pearson correlation coefficient ρ are provided.

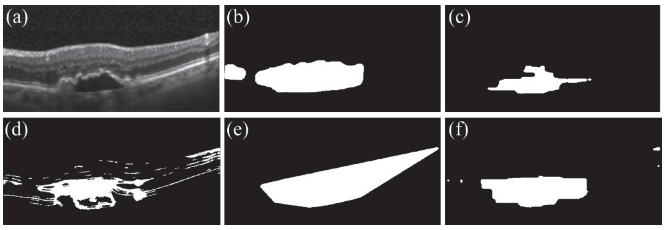


Fig. 7. Qualitative results of the ablation studies, showing anomaly segmentation results on an exemplary sample. (a) Original B-scan, (b) Manual annotation, Segmentation results of (c) WeakAnD (binary), (d) WeakAnD (thresholding), (e) WeakAnD (convex-hull) and (f) WeakAnD.

V. Discussion

We propose to detect and segment anomalies in retinal OCT images using epistemic uncertainty estimations. The approach is built on the assumption that epistemic uncertainty correlates with unknown anatomical variability (anomalies), not present in the training data. This claim is supported by the results, in particular by the high correlation (ρ=0.91) between the amount of anomalous area and uncertainty, as observed in Fig. 6. Another alternative to identify anomalies is to use the entropy of the soft predictions of the layer segmentation method. Using the soft predictions of neural networks directly has been previously explored as an alternative to identify out-of-distribution samples [45], [46]. We used this idea as a baseline to compare with and we observed that epistemic uncertainties are more powerful to reflect abnormal changes with respect to the training set (Table II, Fig. 8). We believe this is caused by the softmax predictions capturing different information than uncertainty estimates obtained through MC-sampling. The soft predictions indicate the probability of a given pixel belonging to a specific class, while an uncertainty estimate provides information regarding the confidence of the network about assigning a specific likelihood. As pointed out by [18], a model can be uncertain in its predictions even with a high softmax output for a specific class.

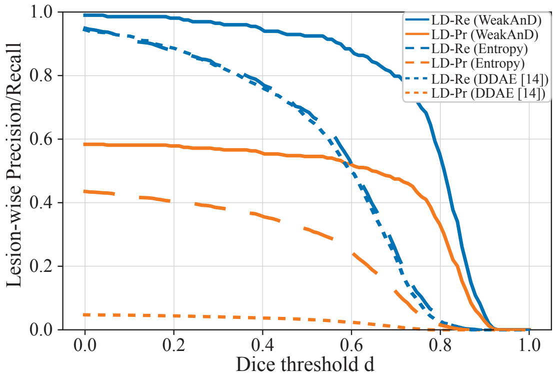


Fig. 8. Lesion detection Recall (LD-Re) and Precision (LD-Pr) curves for the proposed approach (solid) and the baseline methods. The low LD-Precision curve of [14] can be explained by its noisy segmentation results which lead to several tiny false positive lesions.

We also took advantage of weak supervision by training our segmentation model with labels provided by an existing automated approach [24], which is known to perform accurately in healthy scans. Thus, instead of relying on a large training set of normal and diseased patients with costly per-pixel manual

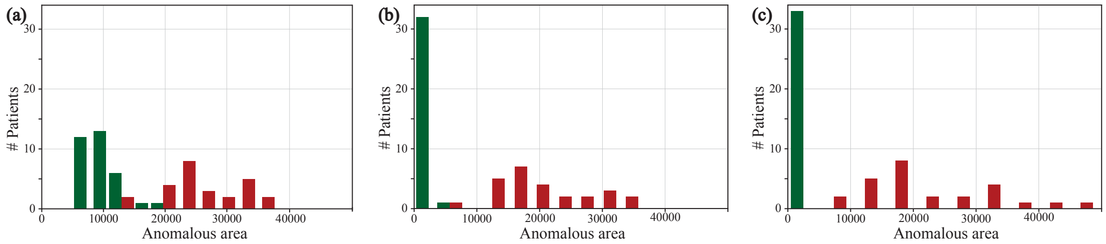


Fig. 9. Histograms of the volume-wise classification experiment for detecting late wet AMD cases. Results of (a) DDAE [14], (b) entropy of soft predictions and (c) our method. The horizontal axis represents the average number of anomalous pixels per B-Scan for each volume and the vertical axis indicates the number of patients. Green and red denote patients from the *normal evaluation* and the *late wet AMD test* datasets, respectively. Note that a separate overview of the classification results of our method is plotted in Fig. 10

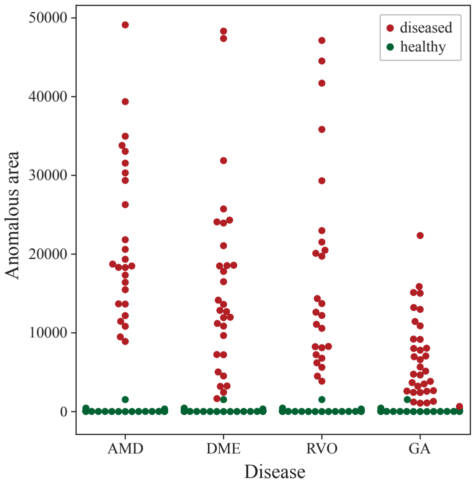


Fig. 10. Categorical scatter plot showing the results of all volume-wise classification experiments (AMD, DME, RVO and GA). Each dot represents a patient volume. Diseases are indicated in the horizontal axis, while the vertical axis represents the average number of anomalous pixels per B-Scan for each volume. Green and red denote patients from the *normal evaluation* and the *diseased* datasets, respectively.

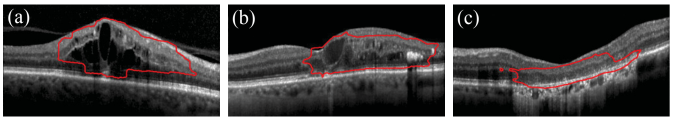


Fig. 11. Qualitative results of the proposed method on (a) DME, (b) RVO and (c) GA cases.

annotations of anomalies, we proposed to train our approach using a normal data set, providing anatomical information via weak labels. Our empirical observations showed that using this alternative still results in high performance. Nevertheless, the segmentation U-Net is not limited to be trained with weak labels: we argue that it could also be trained using manual annotations without loss of generality.

Compared to the baseline method for anomaly detection in OCT [14], our approach achieved significantly better results in terms of several quantitative metrics (Table II). Interpreting these pixel-wise quality measures requires taking into consideration that manually annotating anomalies is a difficult task: transitions between healthy and diseased scans are continuous, often unclear, hard to define and exposed to subjective interpretation. Therefore, ensuring exact and consistent ground truth labellings is nearly impossible. The high degree of overlap between the outputs of our model and the manual annotations indicates then that the proposed approach is able to approximate the performance of a human expert. This is also supported by the fact that the worst observed Dice value (0.72) is relatively high. To complement these interpretations, we also evaluated the performance of the proposed approach to detect lesions as such. The evaluation of the lesion-wise detection experiment, depicted in Fig. 8, linked quantitative pixel-based evaluation metrics with lesion-level detection capabilities. It can be seen that increasing the requirement of Dice performance d for lesions from 0.0 to 0.6 only decreases the lesion detection performance by 10%, as measured in terms of lesion detection precision and recall. These results indicate that the ability of the method to accurately identify the borders of the anomalies does not have a significant effect in the lesion detection performance, as most of the overlap area with the human expert annotation is located in affected tissue. In other words, most of the changes in Dice are explained by differences in the borders of the anomalous regions (as seen in Fig. 5, right column).

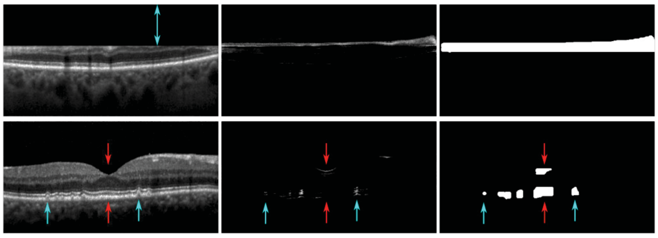


Fig. 12. Anomaly detection in normal scans. Two B-scans from the *normal evaluation* dataset with the largest anomalous area are shown. From left to right: Original B-scans, uncertainty maps and corresponding anomaly segmentation results. Top row: cut edge artifact (blue arrow). Bottom row: small drusen (blue arrows) and false positives (red arrows).

Moreover, it was observed that the size of the predicted anomalous areas was an almost perfect discriminator to classify normal vs. diseased subjects. These volume-wise classification experiments form a proof-of-concept evaluation for a potential screening application in multiple diseases such as wet AMD, dry GA, DME or RVO (Fig. 10). We hypothesize that this is a consequence of our method being able to detect abnormalities in diseased subject without oversegmenting false positives in healthy subjects (Fig. 12).

Although our method does not rely on ground truth annotations of abnormalities for training, it still has hyperparameters that need to be optimized. In our experiments, we used a late wet AMD validation set comprising 5 OCT volumes. We observed that the method is not too sensitive to changes in the dropout rate (Table I). Furthermore, our classification results in different diseases indicate that using a validation set with only one specific condition might be enough to ensure good generalization.

Finally, our approach also reported a significantly better performance than the DDAE method [14] when evaluated both on a lesion and a volume basis. This might be a consequence of the noise in the segmentations generated by the baseline approach, which results in several small isolated regions that increase the false positive rate. Nevertheless, it is important to note that [14] tackles a more difficult task, not only to segment anomalies but to identify categories of them. This restrained such an approach to a localized representation that enables subsequent clustering on local level. In contrast, our method only aims to provide accurate pixel-wise segmentation for the detected abnormalities, not specifically designed for a subsequent clustering step of anomalies. In addition, the method in [14] does not incorporate anatomical information during training nor post processing.

The qualitative analysis of the results in Fig. 5 revealed that the uncertainty maps showed high values surrounding subretinal fluid (SRF), a concave form in cases of pigment epithelial detachments (PED) and dense patterns in regions of hyperreflective foci (HRF). In general, the segmentation model predicted the background class with high confidence in large areas of fluid, probably due to missing edges and/or dark appearance in those regions. This observation highlights the necessity of appropriate post-processing to obtain smooth segmentation maps (Table III).

From the ablation study is also possible to conclude that each part of the method is important to ensure accuracy and consistent results. In particular, we observed that using less informative target labels for the segmentation approach e.g., by targeting the whole retina instead of its constitutive layers (*WeakAnD (binary)* in Table III) decreases the performance for anomaly detection (Dice index drop of 14.7%). We observed that the uncertainty maps produced by the binary alternative were not as detailed and dense as the ones of the proposed method. This caused segmentation shapes inconsistent with the manual anomaly annotations (see Fig. 7(c)), as well as apparent horizontal and vertical gap-artifacts of segmentation areas. Considering the fact that the cellular components of the retina are arranged in a layer-wise manner [47] and pathologies alter their appearance, using retinal layer information proved to be a particularly appropriate way to incorporate anatomical knowledge into the model. This helped to achieve more representative uncertainty values and, therefore, better results. For this particular point, it is important to emphasize the contribution of the post-processing method based on majority-ray-casting. As observed in Table III, replacing this stage by other alternative approaches caused drops in performance. Removing majority-ray-casting and only conducting thresholding of the uncertainty maps (*WeakAnD (thresholding)*) resulted in poor quantitative results, decreasing the Dice index by 28.4%. This is also reflected in Fig. 7(d), where the exemplary segmentation covers not only the anomalous regions but also some borders between retinal layers. This result was obtained using an optimal threshold ($t = 0.03$) selected on the validation set. Although this might compensate for the discontinuous property in the area with true positive anomalies, it brings further layer interfaces to the final segmentation, where a certain degree of uncertainty is also present. Complementing thresholding with a convex-hull based post-processing also caused unwanted artifacts, e.g. in Fig. 7(e), where a small blob in the top right (remaining after thresholding) caused a peculiar segmentation. This is a consequence of the inability of the convex-hull approach to handle multiple non-connected anomalous areas by definition. On the contrary, the anomalous area is better captured when applying our majority-ray-casting method. This indicates the potential of using a relatively straightforward approach combined with an appropriate post-processing step in the context of anomaly detection. Notice that this technique targets a blob-shaped segmentation instead of a specific disease appearance. Our post-processing approach is intended to help to transfer the layered output of the uncertainty estimates to a blob-shaped segmentation surrounding abnormalities, which we believe is the most straightforward way to highlight them in general. The previously published approach is already able to retrieve such a shape (Fig. 3 in [14]), although with significant false positive detections. Our thresholded uncertainty maps, on the other hand, slightly outperform [14] in terms of Dice, but are not able to retrieve such a blob-shape due to the partial blindness of the uncertainty estimates. In other words, majority-ray-casting helps to transfer the layered output of the uncertainty estimates to a continuous representation that delivers an easier-to-interpret result. This is line with what can be seen from Table II and Table III, where [14] reported lower precision but higher recall than our method. Finally it is worth mentioning that, in addition to fluid related lesions (Fig. 5 and Fig. 11 (a)), our approach detects other anomalies such as drusen (Fig. 12, bottom row), hyperreflective material (Fig. 11 (b)), DRIL or GA lesions (Fig. 11 (c)). This demonstrates that the presented method allows to highlight a variety of retinal abnormalities in multiple diseases.

We observed that the network detected anomalies only in the area ranging from the top of the NFL to the RPE. We believe that this is a consequence of the model being restricted by the anatomy used for training. A similar behavior was observed before in the binary model, trained to segment the retina and the background. By using the weak labels generated using the Garvin *et al.* [24] method, our network is unable to capture representative uncertainty estimates in regions that are jointly labeled as background. Our hypothesis is that the network optimizes its loss function by focusing more on the non-background layer labels. This makes the network invariant to changes in areas below the RPE and above the vitreous-macular interface, and therefore does not show uncertainties there. Incorporating labels for other layers such as the choroid might allow the model to explicitly learn the normal characteristics of these regions, and thus show higher uncertainty estimates when deviations from this appearance are observed (e.g. due to hypertransmision).

Finally, it is worth pointing out a potential limitation of the majority-ray-casting algorithm, related to the internal distribution and localization of anomalies within the retina. Since the post-processing algorithm assumes that areas surrounded by uncertainties are anomalous (Fig. 2), there could be specific clinical scenarios in which this assumption does not hold: e.g. in between three independent anomaly detections (Fig. 12, bottom row, bottom red arrow). Hence, this can lead to oversegmentation. In some cases, we also observed false positives in the fovea depression, caused by a thinning in the top retinal layers (Fig. 12, bottom row, top red arrow). Nevertheless, anomaly detection approaches are needed to reach high levels of sensitivity when applied for screening or detecting pathological areas, and false positives are tolerated to a certain extent. Therefore, oversegmentation might not harm the final application. Moreover, the volume-wise disease detection experiment showed perfect separation between diseased and healthy subjects using only the amount of abnormal area for discrimination.

VI. Conclusion

We proposed a weakly supervised anomaly detection method based on epistemic uncertainty estimates from a Bayesian multiclass U-Net model, with application in retinal OCT analysis. The segmentation approach was trained on a cohort of normal subjects to characterize healthy retinal anatomy. No annotations of the target class (anomalies) were used to learn that model. Instead, we took advantage of the fact that traditional segmentation methods work accurately in well-defined environments such as healthy populations, allowing to easily obtain large amounts of segmented data. Following this perspective, we used an automated method [24] to generate weak labels for the individual retinal layers. During test time, unseen B-scans were processed by the Bayesian network, and Monte Carlo sampling with dropout was used to retrieve epistemic uncertainty estimates. To better exploit its application to segment potential anomalies, a novel post-processing technique based on *majority-ray-casting* was introduced. The final output was a binary mask with smooth segmentation of retinal abnormalities.

The proposed anomaly detection approach needs only healthy samples for training, detects the deviation from normal by exploiting the injected anatomical information of healthy scans and is therefore–by definition–not limited to a specific disease or pathology. An extensive evaluation using 33 normal and 115 diseased OCT volumes (1617 and 5635 B-scans, respectively) demonstrates that our uncertainty-driven method is able to detect anomalies under several conditions, outperforming alternative approaches. This makes it a promising tool in the context of biomarker discovery, where the detection and exploration of atypical visual variability is a fundamental task. In this context, further research is planned to explore the suitability of the presented method in the context of biomarker detection.

References

[1] S. B. Nimse, M. D. Sonawane, K.-S. Song, and T. Kim, “Biomarker detection technologies and future directions,” *Analyst*, vol. 141, no. 3, pp. 740–755, 2016.

[2] A. J. Asman and B. A. Landman, “Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE),” *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1779–1794, 2011.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. 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, 2017.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.

[5] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, “Large scale deep learning for computer aided detection of mammographic lesions,” *Medical Image Analysis*, vol. 35, pp. 303–312, 2017.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *arXiv preprint arXiv:1711.05225*, 2017.

[7] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. ODonoghue, D. Visentin *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nature Medicine*, vol. 24, no. 9, p. 1342, 2018.

[8] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, “Artificial intelligence in retina,” *Progress in Retinal and Eye Research*, 2018.

[9] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, “Fully automated detection and quantification of macular fluid in OCT using deep learning,” *Ophthalmology*, vol. 125, no. 4, pp. 549–558, 2018.

[10] H. Lee, K. E. Kang, H. Chung, and H. C. Kim, “Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration,” *American journal of ophthalmology*, vol. 191, pp. 64–75, 2018.

[11] S. G. Zadeh, M. W. Wintergerst, V. Wiens, S. Thiele, F. G. Holz, R. P. Finger, and T. Schultz, “Cnns enable accurate and fast segmentation of drusen in optical coherence tomography,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 65–73.

[12] T. Schlegl, H. Bogunovic, S. Klimscha, P. Seeböck, A. Sadeghipour, B. Gerendas, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, “Fully automated segmentation of hyperreflective foci in optical coherence tomography images,” *arXiv preprint arXiv:1805.03278*, 2018.

[13] J. I. Orlando, P. Seeböck, H. Bogunović, S. Klimscha, C. Grechenig, S. Waldstein, B. S. Gerendas, and U. Schmidt-Erfurth, “U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans,” *arXiv preprint arXiv:1901.07929*, 2019.

[14] P. Seeböck, S. Waldstein, S. Klimscha, H. Bogunovic, , T. Schlegl, B. S. Gerendas, R. Donner, U. Schmidt-Erfurth, and G. Langs, “Unsupervised identification of disease marker candidates in retinal OCT imaging data,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1037–1047, 2019.

[15] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014.

[16] “Towards trustable machine learning,” *Nature Biomedical Engineering - Editorial*, vol. 2, no. 10, pp. 709–710, 2018. [Online]. Available: https://doi.org/10.1038/s41551-018-0315-x

[17] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[18] Y. Gal and Z. Ghahramani, “Bayesian convolutional neural networks with Bernoulli approximate variational inference,” *arXiv preprint arXiv:1506.02158*, 2015.

[19] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” *arXiv preprint arXiv:1511.02680*, 2015.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017, pp. 5574–5584.

[21] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, p. 17816, 2017.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*. Springer, 2015, pp. 234–241.

[23] K. A. Joltikov, C. A. Sesi, V. M. de Castro, J. R. Davila, R. Anand, S. M. Khan, N. Farbman, G. R. Jackson, C. A. Johnson, and T. W. Gardner, "Disorganization of Retinal Inner Layers (DRIL) and neuroretinal dysfunction in early diabetic retinopathy," *Investigative Ophthalmology & Visual Science*, vol. 59, no. 13, pp. 5481–5486, 2018.

[24] M. K. Garvin, M. D. Abràmoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 9, pp. 1436–1447, 2009.

[25] J. Fujimoto and E. Swanson, "The development, commercialization, and impact of optical coherence tomography," *Investigative Ophthalmology and Visual Science*, vol. 57, no. 9, 2016.

[26] 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," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[27] J. Jonas, M. Paques, J. Monés, and A. Glacet-Bernard, "Retinal vein occlusions," in *Macular Edema*. Karger Publishers, 2010, vol. 47, pp. 111–135.

[28] G. S. Tan, N. Cheung, R. Sim, G. C. M. Cheung, and T. Y. Wong, "Diabetic macular oedema," *The Lancet Diabetes & Endocrinology*, vol. 5, no. 2, pp. 143–155, 2017.

[29] W.-D. Vogl, S. M. Waldstein, B. S. Gerendas, T. Schlegl, G. Langs, and U. Schmidt-Erfurth, "Analyzing and predicting visual acuity outcomes of anti-VEGF therapy by a longitudinal mixed effects model of imaging and clinical data," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 10, pp. 4173–4181, 2017.

[30] U. Schmidt-Erfurth and S. M. Waldstein, "A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration," *Progress in Retinal and Eye Research*, vol. 50, pp. 1–24, 2016.

[31] F. L. Ferris, C. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, and S. R. Sadda, "Clinical classification of age-related macular degeneration," *Ophthalmology*, vol. 120, no. 4, pp. 844–851, 2013.

[32] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, p. 158, 2018.

[33] J. I. Orlando, J. Barbosa Breda, K. van Keer, M. B. Blaschko, P. J. Blanco, and C. A. Bulant, "Towards a glaucoma risk index based on simulated hemodynamics from fundus images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Springer International Publishing, 2018, pp. 65–73.

[34] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 655–663.

[35] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[36] N. Pawlowski, M. C. Lee, M. Rajchl, S. McDonagh, E. Ferrante, K. Kamnitsas, S. Cooke, S. Stevenson, A. Khetani, T. Newman *et al.*, "Unsupervised lesion detection in brain CT using bayesian convolutional autoencoders," in *Medical Imaging with Deep Learning*, 2018.

[37] P. A. Dufour, H. Abdillahi, L. Ceklic, U. Wolf-Schnurrbusch, and J. Kowal, "Pathology hinting as the combination of automatic segmentation with a statistical shape model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 599–606.

[38] D. Sidibé, S. Sankar, G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, G. S. Tan, D. Milea, E. Lamoureux, T. Y. Wong *et al.*, "An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images," *Computer Methods and Programs in Biomedicine*, vol. 139, pp. 109–117, 2017.

[39] S. Sedai, B. Antony, D. Mahapatra, and R. Garnavi, "Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning," in *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 2018, pp. 219–227.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[41] W.-D. Vogl, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and G. Langs, "Predicting macular edema recurrence from spatio-temporal signatures in optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1773–1783, 2017.

[42] H. Bogunović, A. Montuoro, M. Baratsits, M. G. Karantonis, S. M. Waldstein, F. Schlanitz, and U. Schmidt-Erfurth, "Machine learning of the progression of intermediate age-related macular degeneration based on OCT imaging," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 6, pp. BIO141–BIO150, 2017.

[43] K. He, X. Zhang, S. Ren *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. of IEEE ICCV*, 2015, pp. 1026–1034.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

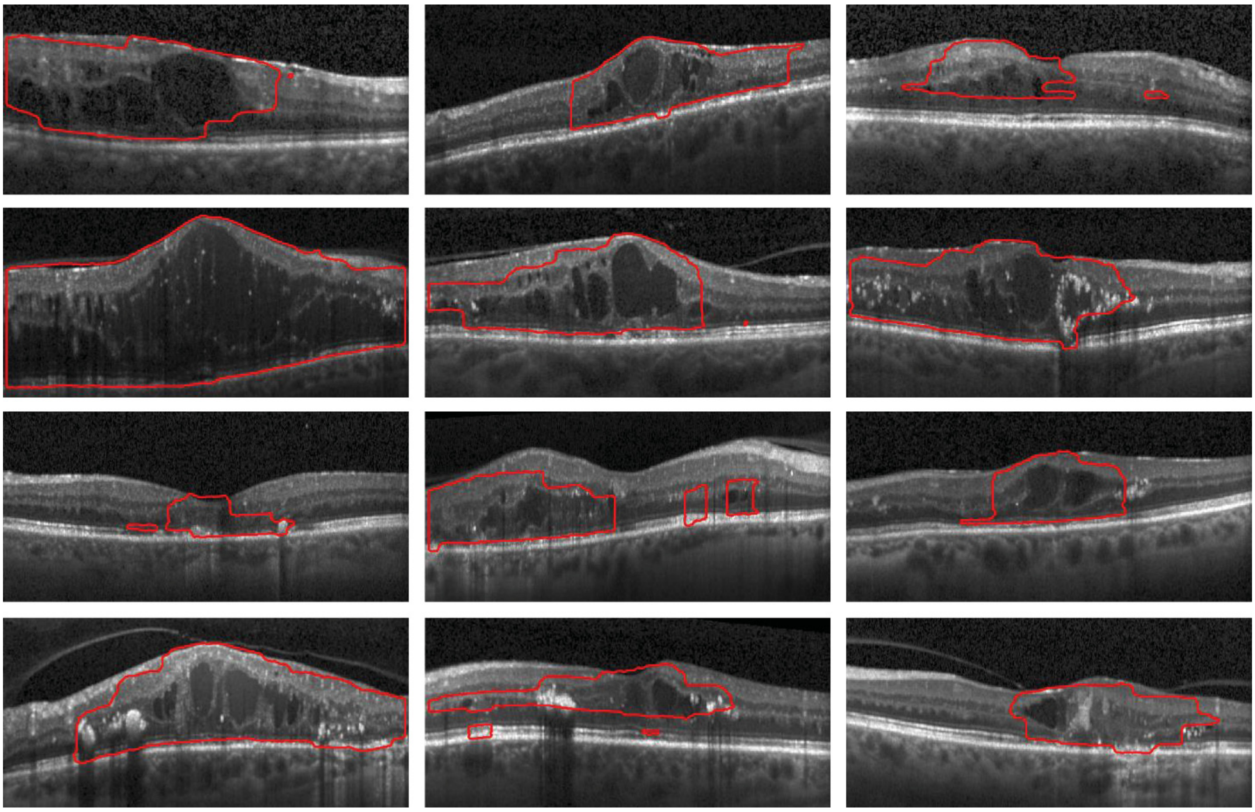
[45] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[46] N. J. Ratzlaff, "Methods for detection and recovery of out-of-distribution examples," 2018.

[47] G. D. Hildebrand and A. R. Fielder, "Anatomy and physiology of the retina," in *Pediatric Retina*. Springer, 2011, pp. 39–65.

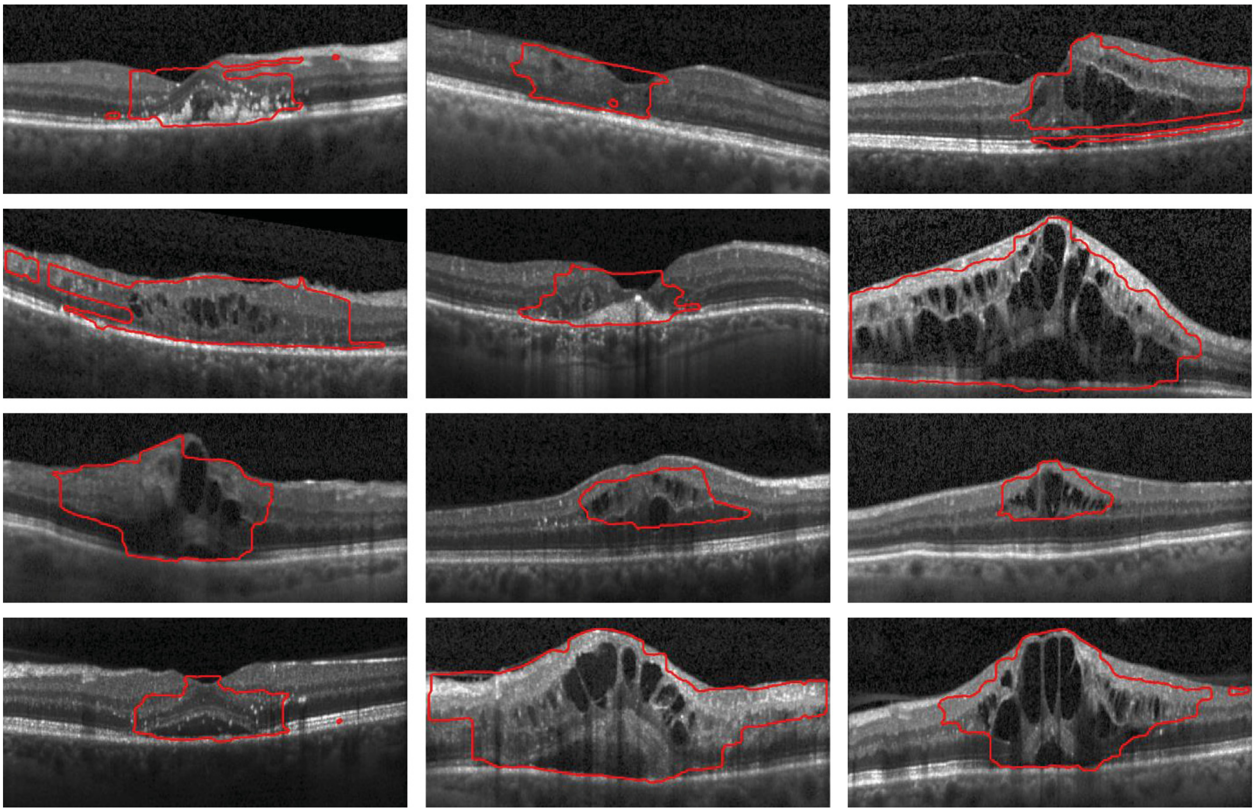
Supplementary Material

S1. Qualitative Results – DME



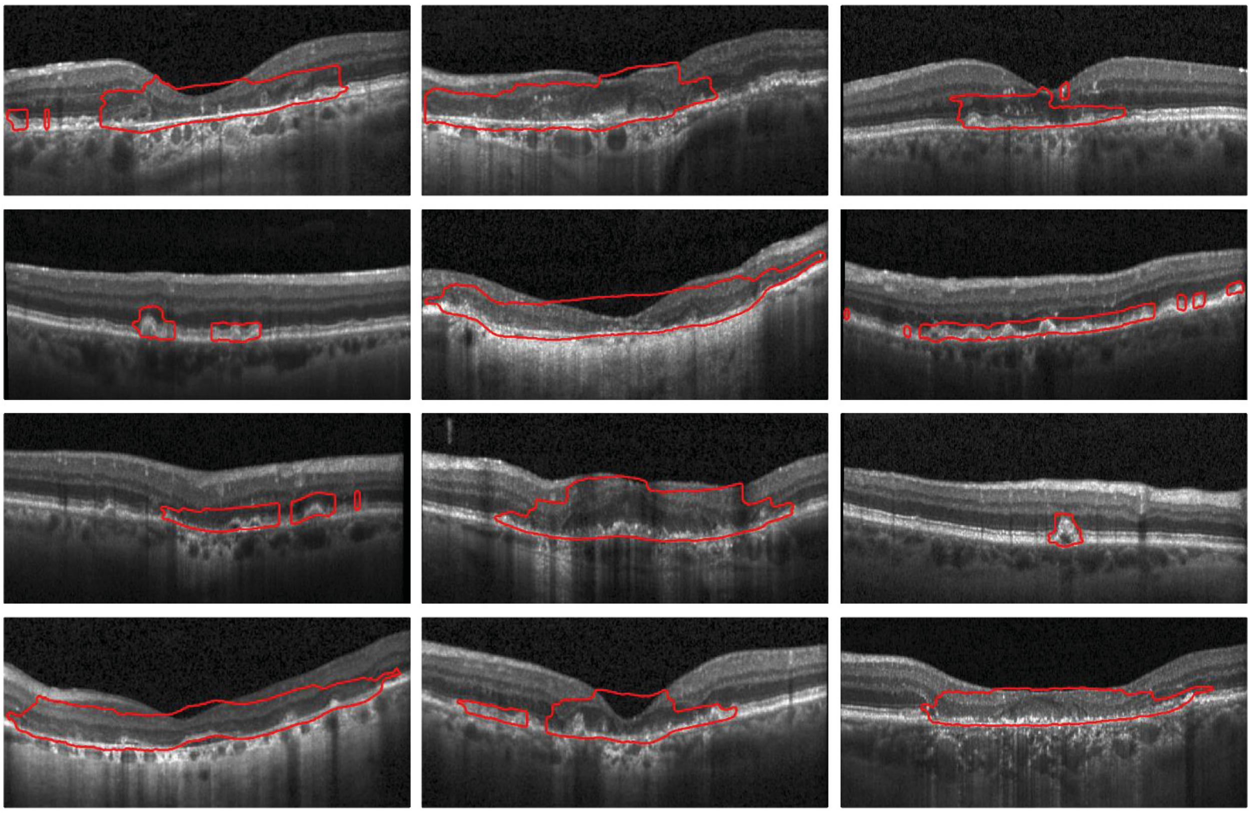
Supplementary Figure 1. Qualitative results of the proposed anomaly detection method on the diabetic macular edema (DME) dataset. All images correspond to central B-scans of the volumes. The red contours depict the boundaries of the segmented anomalous regions.

S2. Qualitative Results – RVO



Supplementary Figure 2. Qualitative results of the proposed anomaly detection method on the retinal vein occlusion (RVO) dataset. All images correspond to central B-scans of the volumes. The red contours depict the boundaries of the segmented anomalous regions.

S3. Qualitative Results – GA



Supplementary Figure 3. Qualitative results of the proposed anomaly detection method on the geographic atrophy (GA) dataset. All images correspond to central B-scans of the volumes. The red contours depict the boundaries of the segmented anomalous regions.

CHAPTER 7

A Paradigm Shift in Retinal Biomarker Identification by Unsupervised Deep Learning

"Investigate what is
and not what pleases."

– Johann Wolfgang von Goethe

In the following we perform biomarker discovery based on an *unsupervised feature learning* strategy. The aim of the method is to capture phenotype characteristics of a diseased patient population, without the need of manual annotations for training. This means that large-scale image data is explored in an unbiased way, remaining invariant to established medical hypotheses.

We propose to learn disease specific features from OCT images in a completely unsupervised manner, utilizing a multi-stage autoencoder based approach. In this way, both a representation of the local morphology of the retina and the whole volume is learned, allowing an analysis on different levels. First, we use a deep autoencoder to learn a low-dimensional embedding of A-scans. This feature representation encodes *local* characteristics of the retina, representing the morphological information at a specific position across all retinal layers. In a second stage, a convolutional autoencoder is trained on the local embeddings of the first stage, learning a low-dimensional *global* volume representation. We extensively evaluate the learned features to asses their potential as biomarkers. Besides a qualitative analysis, we correlated the features with morphological attributes of retinal morphology as conventionally measured from OCT, investigated their relation to measures of disease activity obtained by fluorescein angiography and evaluated their ability to predict visual function.

The following manuscript *”A paradigm shift in retinal biomarker identification by unsupervised deep learning”* is planned for submission to a journal.

A paradigm shift in retinal biomarker identification by unsupervised deep learning

Philipp Seeböck[a,b,†], Sebastian M. Waldstein[a,†], Rene'e Donner[a,b], Amir Sadeghipour[a], Hrvoje Bogunovic[a], Aron Osborne[c], Georg Langs[b], Ursula Schmidt-Erfurth[a]*

[a]*Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology and Optometry, Medical University Vienna, Austria.*

[b]*Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Austria*

[c]*Genentech, Inc, 1 DNA Way, South San Francisco, CA, United States.*

Abstract

Manageable and clinically meaningful biomarkers are an enormous unmet need for personalized health care using diagnostic imaging data. High-resolution retinal imaging by optical coherence tomography offers abundant individualized medical data and the opportunity for deep phenotyping. Artificial intelligence has successfully been leveraged to evaluate specifically defined retinal biomarkers. However, this requires prohibitively large amounts of human annotations of training data, and inherently carries a bias caused by the fallible and dogmatic nature of human experts who predefine the biomarkers. Here, our aim was to create an unbiased vocabulary of biomarkers representing the most important patterns of light-tissue interaction in three-dimensional retinal imaging using optical coherence tomography. We introduce a rigorously data-driven unsupervised deep learning approach, resulting in identification of biomarker candidates without any restricting input or domain knowledge beyond raw images. We demonstrate that the identified features correlate well with specific biomarkers traditionally used in clinical practice, and largely surpass them in the ability to correlate visual function with retinal structure in our population. In addition, our method is also able to discover hitherto unknown biomarker candidates. The complex morphologic information contained in retinal imaging is condensed into an extremely compact representation of only 20 features, offering a perspective on how big data in the medical imaging domain can efficiently be made manageable in the near future.

Keywords: unsupervised learning, optical coherence tomography, biomarker discovery, autoencoder, age-related macular degeneration

1. Introduction

Personalized health care based on big -omics data requires efficient biomarkers as input for predictive models. Ideally, these biomarkers would be characteristic of the data, unbiased, compact, task-independent, and easy to obtain. In the field of ophthalmology, modern retinal imaging by optical coherence tomography (OCT) shows great potential for personalized health care applications. OCT is an affordable, non-invasive imaging technique that acquires huge datasets of high-resolution three-dimensional images within instants [9]. It has become the most important diagnostic test in ophthalmology, with approximately 30 million procedures annually or an OCT scan taken every few seconds worldwide [6]. This leaves researchers and practitioners overwhelmed by millions of images and a lacking consensus regarding the relevant imaging biomarkers for an efficient management of the leading diseases of modern times such as diabetic retinopathy and age-related macular degeneration.

The automated analysis of retinal imaging data by artificial intelligence is currently evolving as a paradigm-shifting tool to bridge the gap between rapid advances in imaging hardware and the challenges in big-data analysis [17]. A major issue is that native OCT volumes are too high-dimensional to be effectively used as an input for predictive models. Thus, specific pre-defined biomarkers such as retinal fluid typically serve as input for subsequent analysis steps [13, 16]. Recent pioneering studies have demonstrated that these established biomarkers can be reliably assessed and measured by supervised deep learning, a group of computer methods where an algorithm learns to replicate human behavior by

*corresponding author. *email:* ursula.schmidt-erfurth@meduniwien.ac.at using prior experience in the form of tens of thousands of manually labelled samples [14, 5]. This approach has critical disadvantages because it can only find what is defined a-priori by human experts, thus being limited to known biomarkers, and it scales poorly due to the need for labor-intensive ground-truth data.

To surpass these limitations, we introduce a novel concept of exploring complex retinal imaging data. Instead of manual labeling and supervised learning, we propose unsupervised deep learning for unbiased feature generation [15]. We introduce an artificial intelligence algorithm that teaches itself to capture the most characteristic local structural markers in OCT images, which represent the main patterns of light-tissue interaction as OCT is acquired. In a second step, we achieve an extremely compact global description of the impractically complex three-dimensional OCT scans. To validate our method, we demonstrate that the obtained features correlate better with clinically relevant measures of retinal function and disease activity than established biomarkers quantified by conventional supervised deep learning.

2. Results

2.1. Deep learning pipeline

The proposed artificial intelligence pipeline (Figure 2) consists of two auto-encoders to capture (a) the most important local features in the 3D image stack; and (b) a compact global description of the features obtained in the previous step.

In principle, an auto-encoder comprises two sequential deep neural networks. The first (encoding) network is trained to produce high-level low-dimensional descriptors of input data (e.g., an image), while the second (decoding) network is trained to reconstruct the original input data from the high-level description provided by the encoding network. If the reconstruction is accurate (i.e. if the output of the decoder matches the input of the encoder), we can assume that a meaningful high-level representation (or embedding) of the input data has been learned. These learned features serve as novel biomarker candidates in our experiments.

OCT images are acquired by scanning a laser beam tomographically across the retina and sampling the light-tissue interaction at each individual scanning location. Thus, we applied the first auto-encoder on these individual scanning locations resembling vertical signal columns (A-scans, 1×1×1024) to learn a 20-dimensional embedding of the local light-tissue interaction. The activation of the 20 learned local features can be displayed and interpreted as feature maps (Figure 2 and Figure 1). We then applied the second auto-encoder on 3D volumes comprised of the obtained local embeddings (512×128×20) and learned a 20-dimensional embedding of the full volumes (Figure 2). Thus, we receive 20 global volume features that represent the main spectrum of morphologic patterns of a 3D image.

The auto-encoders were trained on a dataset consisting of 54,900 OCT volume scans (512×128×1024) of 1,094 patients enrolled in a classic randomized clinical trial described elsewhere (*NCT00891735*) [1]. To validate our approach, we used the baseline condition, when all patients presented with treatment nave neovascular age-related macular degeneration in the study eye, and evaluated the correspondence of the identified biomarker candidates with clinically established markers. These included markers of visual function (best corrected visual acuity and low luminance visual acuity), markers of retinal morphology as conventionally measured from OCT (retinal thickness, volume of intraretinal and subretinal fluid, volume of pigment epithelial detachment) as well as measures of disease activity obtained by fluorescein angiography, a conventional, invasive dye-based investigation (total area of lesion, total area of leakage) [18]. We did not perform correction for multiplicity for the conducted statistical tests, due to the explorative nature of this work. This procedure was selected in order not to increase the type II error in the biomarker search.

2.2. Local features

The 20 learned unsupervised local features (a1 - a20) captured the local morphologic patterns in the OCT data to a high degree and corresponded well to conventional OCT features, but also provided previously unknown features, i.e. biomarker candidates that had not been considered yet in clinical practice (Figure 1). The most relevant features are analyzed in detail below. Univariate correlations between the average activation of the individual features per OCT volume and the validation metadata are presented in Table 1. In general, correlations were stronger for anatomical metadata (r up to 0.73) than for functional metadata (r up to -0.40).

Machine learning regression was performed to evaluate the capability of all combined local features to represent retinal morphology, visual function and disease activity. The prediction of the model for functional and anatomical metadata is shown in Table 2.

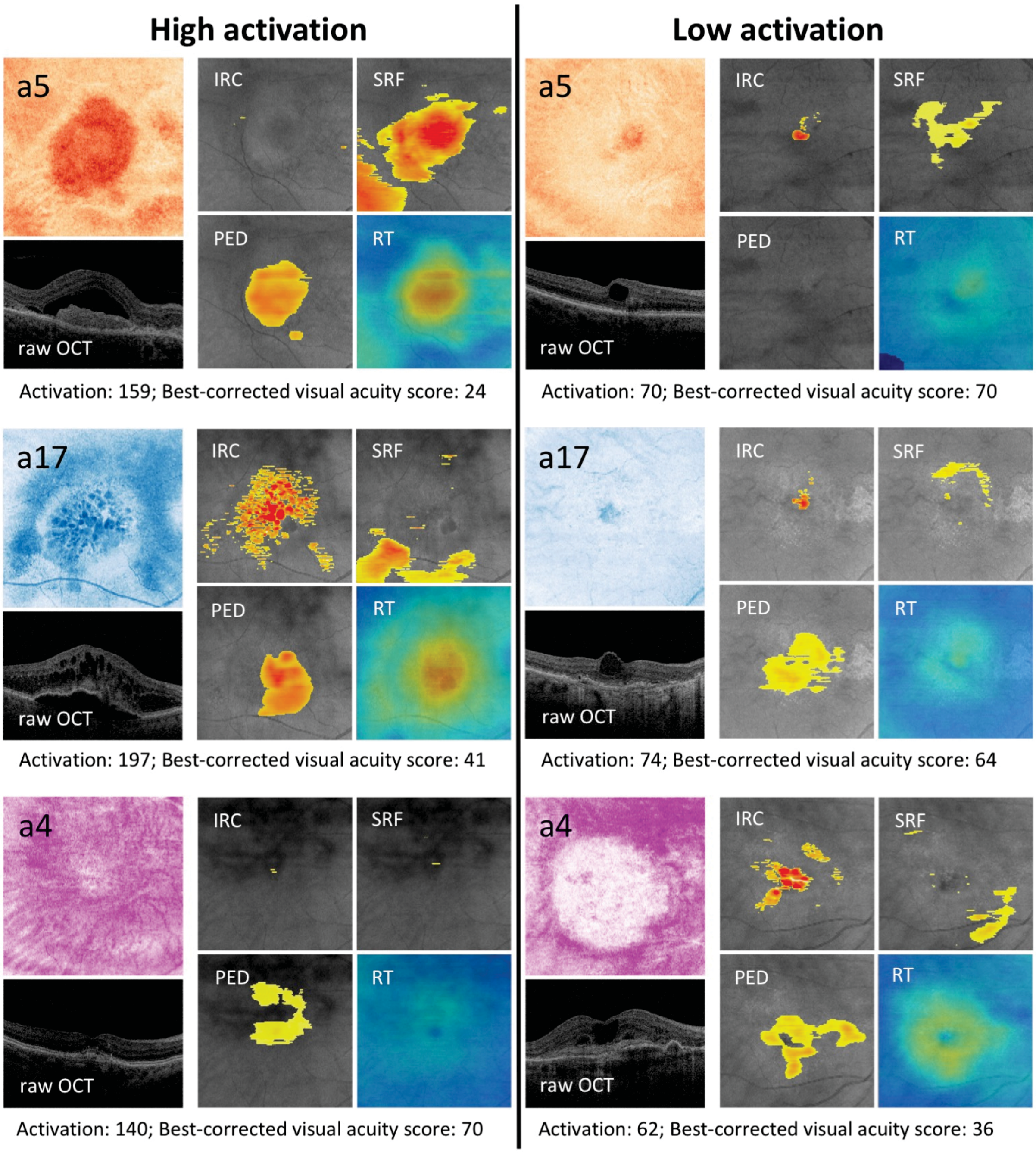

High activation
Low activation
a5
IRC
SRF
PED
RT
raw OCT
Activation: 159; Best-corrected visual acuity score: 24
a5
IRC
SRF
PED
RT
raw OCT
Activation: 70; Best-corrected visual acuity score: 70
a17
IRC
SRF
PED
RT
raw OCT
Activation: 197; Best-corrected visual acuity score: 41
a17
IRC
SRF
PED
RT
raw OCT
Activation: 74; Best-corrected visual acuity score: 64
a4
IRC
SRF
PED
RT
raw OCT
Activation: 140; Best-corrected visual acuity score: 70
a4
IRC
SRF
PED
RT
raw OCT
Activation: 62; Best-corrected visual acuity score: 36

Figure 1: Representative examples of feature maps obtained by the local embedding. The composites to the right of each column show heatmaps of conventional biomarkers obtained by validated segmentation algorithms using supervised deep learning. High and low activation of the detected new biomarkers with concomitant visual function are shown side-by-side. *Top row:* Feature a5 demonstrates a pronounced negative structure-function correlation, despite a low correspondence to retinal fluid, which is the conventional marker attributed a high relevance for vision. We assume that this biomarker candidate corresponds to subretinal hyperreflective material (arrow). *Middle row:* Feature a17 demonstrates the best correlation with markers of exudation as conventionally measured in OCT. An excellent correspondence is for instance observed for intraretinal cystoid fluid (compare the lobulated pattern). *Bottom row:* Feature a4 represents a new biomarker candidate discovered in this work. The marker does not intrinsically correspond to any known clinical entity in OCT images. Remarkably, a positive correlation between the activation of a4 and visual function markers was noted. IRC = intraretinal cystoid fluid; PED = pigment epithelial detachment; RT= retinal thickness; SRF = subretinal fluid.

	Functional		OCT				Fluorescein Angiography	
	Best Corrected Visual Acuity	Low Luminance Visual Acuity	Retinal Thickness	Intra-Retinal Cystoid Fluid	Subretinal Fluid	Pigment Epithelial Detachment	Total Area of Lesion	Total Area of Leakage
a1	-0.21	-0.19	0.30	0.07	0.25	0.15	0.17	0.17
a2	-0.20	-0.17	0.36	0.20	0.32	0.15	0.11	0.13
a3	-0.10	-0.08	0.39	0.18	0.27	0.10	0.06	0.08
a4	0.30	0.36	-0.32	-0.07	-0.22	-0.24	**-0.32**	**-0.33**
a5	**-0.31**	**-0.40**	0.30	0.20	0.33	0.19	0.25	0.31
a6	-0.10	-0.03	0.31	0.07	0.19	0.12	0.07	0.05
a7	-0.08	-0.20	-0.05	-0.00	0.17	0.17	0.20	0.19
a8	-0.08	-0.04	0.29	0.14	0.24	0.11	0.06	0.06
a9	-0.17	-0.21	0.27	0.05	0.15	0.09	0.10	0.17
a10	-0.29	-0.27	0.54	0.20	0.41	0.24	0.22	0.24
a11	-0.05	-0.03	0.03	-0.02	0.11	0.07	0.06	0.04
a12	-0.08	0.00	0.49	0.18	0.26	0.10	-0.01	0.02
a13	-0.22	-0.26	0.41	0.27	0.39	0.22	0.18	0.19
a14	0.05	0.03	-0.31	-0.04	-0.04	-0.09	-0.10	-0.10
a15	0.03	0.02	0.01	-0.04	0.08	-0.01	0.08	0.08
a16	-0.02	-0.01	0.08	-0.04	0.16	0.10	0.09	0.05
a17	-0.28	-0.23	**0.73**	**0.37**	**0.45**	**0.32**	0.23	0.26
a18	-0.23	-0.26	0.18	0.00	0.24	0.17	0.22	0.22
a19	-0.19	-0.16	0.28	0.10	0.28	0.15	0.13	0.12
a20	-0.22	-0.19	0.39	0.18	0.33	0.21	0.16	0.17

Table 1: Univariate Pearson correlation coefficients between the 20 identified unsupervised local features (a1-a20) and functional variables as well as measures of disease activity by OCT and fluorescein angiography. Green colour indicates a positive, and blue colour a negative correlation. The level of correlation is colour coded, and the strongest correlation for each variable are shown in boxes. Correlations with no significant difference from 0 are greyed out.

	Visual Function		Optical coherence tomography				Fluorescein angiography	
	BCVA (letter score)	LLVA (letter score)	RT (μm)	IRC (mm^3)	SRF (mm^3)	PED (mm^3)	Lesion area (mm^2)	Leakage area (mm^2)
Local features								
R^2	0.26	0.44	0.65	0.09	0.44	0.20	0.27	0.22
MAE	9.3±7.1	10.3±6.5	10.6±11.0	62±48	333±3.3e6	300±248	1.2±1.0	1.3±1.0
Global features								
R^2	0.29	0.46	0.64	0.19	0.27	0.28	0.21	0.15
MAE	8.9±7.3	9.7±7.0	10.9±11.0	54±50	342±412	286±237	1.4±1.0	1.3±0.8

Table 2: Machine learning prediction of functional and morphologic target variables from local and global features. For each outcome variable, the coefficient of determination (R) and mean absolute error (MAE) are shown. BCVA, best-corrected visual acuity; IRC = intraretinal cystoid fluid; LLVA = low luminance visual acuity; PED = pigment epithelial detachment; RT = retinal thickness; SRF = subretinal fluid.

2.3. Interpretation of selected local features

The features with the largest correlation coefficients for each individual meta-variable are further analyzed below and presented in detail in Figure 1. Feature a5 achieved the best correlation with functional target variables, i.e. best-corrected visual acuity (r = -0.31) and low luminance visual acuity (r = -0.40). Interestingly, a5 did not show strong correlations with the quantified morphologic variables such as retinal thickness or fluid. However, a5 visually corresponded to hyperreflective subretinal lesions that may represent subretinal fibrosis with photoreceptor function loss (Figure 1).

Feature a17 corresponded best to the conventional fluid-related markers in OCT images, including retinal thickness (r = 0.73), intraretinal cystoid fluid volume (r = 0.37), subretinal fluid volume (r = 0.45) and pigment epithelial detachment volume (r = 0.32). Figure 1 illustrates the excellent topographic correspondence between a17 and segmentations of intraretinal cystoid fluid and subretinal fluid. The retinal vasculature was also captured by a17.

Feature a4 demonstrated the highest (negative) correlation with conventional features obtained by fluorescein angiography. A4 also surprisingly revealed a markedly positive correlation with retinal function, i.e. r = 0.30 with best-corrected visual acuity and r = 0.36 with low luminance visual acuity. This marker was negatively correlated with the known OCT markers. Clinically, the feature did not correspond intrinsically to any marker that is currently used in ophthalmic practice. Thus, our deep learning network identified a hitherto unconsidered biomarker candidate with a high relevance for visual function. The marker captured the typical pattern of the large choroidal vasculature, and in patients with lower visual acuity showed central punched out regions of low activation (Figure 1).

2.4. Global features

The second autoencoder provided 20 global features per OCT volume scan (v1 - v20). The univariate correlation of the features with functional and morphologic metadata is shown in supplementary Table S1. The global features do not contain interpretable spatial information; thus, a simple correlation to image structures similar to the local features is not possible. Generally, the univariate correlations between the global features and clinical metadata were slightly less strong compared to the local features.

Again, multivariate regression analysis was carried out to investigate the capability of all global markers combined to represent retinal morphology, visual function and disease activity. Results of the regression analysis are provided in Table 2. In general, the global features captured the variability in the metadata similarly well as the local features; however, while using a much simpler description of the OCT data (20 variables per volume versus 512×128×20 variables).

2.5. Descriptive power of novel unsupervised features versus conventional supervised features

We further evaluated the descriptive power of features obtained by the newly developed image analysis approach against conventional biomarkers obtained by supervised deep learning. For this experiment, we compared the prediction model for visual function based on the new features (Table 2) against a separate prediction model based on traditional markers, i.e. the following variables: Intraretinal cystoid fluid (volume and area), subretinal fluid (volume and area), pigment epithelial detachment (volume and area) and mean retinal thickness. Each of these features was quantified in the central 1mm cylinder centered on the fovea centralis, the 1-3mm ring and the area outside the 3mm ring, resulting in 21 variables for a fair comparison to the 20 unsupervised variables. Using these conventional variables, the coefficients of determination were R^2 = 0.20 (MAE: 9.3 ± 7.3) for best corrected visual acuity (p = 0.02 against novel global features), and $R^2 = 0.29$ (MAE: 11.5 ± 7.9) for low luminance visual acuity ($p = 0.05$ against novel global features).

3. Discussion

Supervised deep learning based on manually labelled input data can successfully replicate the behavior of human experts in relatively simple, but labor-intensive tasks such as in triaging retinal OCT scans [5]. However, it has critical limitations, including (1) bias introduced by the underlying domain knowledge used to generate the man-made training data, and (2) limited scalability due to often prohibitively large amounts of annotated data required. These limits of supervised deep learning have been elegantly circumvented by reinforcement learning, where the computer program AlphaGo Zero achieved superhuman performance in playing the game Go by solely being taught the game rules [20]. In medical imaging however, diagnostic procedures and decisions are not nearly as clear as the rules of a board game, and novel innovative approaches are required particularly as therapeutic implications are often controversial and real world outcomes generally poor. By introducing unsupervised deep learning to retinal image analysis, we create a rigorously data-driven analytical tool that is (1) unbiased because it does not rely on human-defined features or hypotheses, and (2) scalable at will because it does not require annotated training data. Our unsupervised deep learning pipeline identified clinically relevant biomarker candidates in a large-scale OCT dataset that were as good as, or better, in representing the visual acuity of patients in a large patient cohort than conventional manually defined features measured by state-of-the-art supervised deep learning methods. We believe our method produces biomarkers that are characteristic of the data, unbiased, compact, task-independent, and easy to obtain.

One major advantage of unsupervised learning is that it automatically learns the most characteristic image features in a dataset, while remaining invariant to any prefabricated and hence biased medical hypotheses. In our experiments, the deep learning algorithm captured the main local biomarkers conventionally used in OCT interpretation, including retinal thickness, intraretinal cystoid fluid, subretinal fluid and pigment epithelial detachment [18]. In addition, it recognized subretinal hyperreflective lesions (a5), which are thought to represent incipient fibrosis, as an important feature unrelated to exudation [22]. In fact, this particular feature showed the strongest correlation with visual function in our cohort; albeit currently not being considered as an endpoint in trials for retinal therapeutics. Thus, our disruptive approach of an unbiased biomarker search may be useful in identifying, defining and prioritizing targets and endpoints for the development of new compounds and interventions. In addition to representing the main known characteristics, our method may also be used to discover new marker candidates in image data. To complement the usual suspects, the deep learning algorithm identified a new feature (a4), which demonstrated a pronouncedly positive correlation with visual acuity in our cohort of patients. Further research may be directed at identifying anatomical correlates for this marker, such as intact neurosensory structures which do not attract any attention in current clinical trials despite morphological appearance. Currently, we are unable to pinpoint individual local markers to a particular anterior-posterior location in OCT A-scans, and therefore the exact origin of the feature activation is yet unknown. Nevertheless, biomarker discovery such as reported here may become an important aspect in medical image interpretation as conventional markers are regarded to have substantial weaknesses in reflecting visual function and disease activity, and pivotal drug developments fail presumably also due to the lack of reliable endpoints [21, 8].

Obviously, not all local markers represent clinically relevant information similar to the OCT image itself. For instance, feature a11 showed very low correlations with the provided meta-information, while showing homogenous activation patterns across images. We may speculate that some of our features, including a11, capture image characteristics such as noise, that are part of the image, but do not relate to individual biomarkers. Future work may address the automated establishment of the number of dimensions in the auto-encoder embedding and thus analyze how many individual features are required to represent an image dataset comprehensively.

The second step of the unsupervised embedding resulted in a characteristic, compact representation of complex three-dimensional OCT datasets. Without human-made limitation to particular variables or measurements, our algorithm provides 20 quantified global features for each volume that represent the major morphologic patterns in the image data, as opposed to an unmanageable amount of 67 million (512×128×1024) voxels in the native image. Despite this heavy compression, the resulting measurements still correlate well with visual acuity, conventional markers on OCT, and multimodal markers of disease activity (e.g. on fluorescein angiography), and surpass conventional OCT markers obtained by supervised learning in representing visual function. Once validated in prospective studies, we believe that unbiased, manageable descriptions of OCT such as the one presented here may be applied in clinical and research practice because they could significantly facilitate the interpretation of complex imaging data, and make therapeutic decision making based on imaging studies at the same time simpler, as well as more reliable.

In our experiments, we achieved a coefficient of determination of R2 = 0.29 and R2 = 0.46 between unsupervised global features and best-corrected and low-luminance visual acuity, respectively, in a large cohort of patients with neovascular age-related macular degeneration at the native stage. The results on best-corrected visual acuity compare favorably with those previously reported in the literature for large datasets, and were indeed superior to the correlations achieved by using conventional OCT markers such as fluid volume, which highlights the value of our approach [16]. Interestingly, the correlation with low-luminance visual acuity was consistently larger than with best-corrected visual acuity. Previous studies have shown impairment in low-luminance visual acuity in patients with age-related macular degeneration that exceeds the deficits seen in best-corrected visual acuity [7, 12, 4]. Possibly, morphologic changes on OCT correspond better to low-luminance visual acuity because it is a more sensitive measurement of visual dysfunction in the macular area.

Unsupervised deep learning has previously been leveraged in medical imaging. For instance, an unsupervised learned representation of local 2D image patches was used for the task of mammography risk scoring [10]. In the context of ophthalmic imaging, researchers have proposed algorithms to identify abnormal tissue patterns by learning the characteristic appearance of normal tissue [14, 19]. In such an approach, anomalous regions can be analyzed further to establish clusters of biomarkers allowing to define marker categories. In contrast, we propose to learn both local and global high-level descriptions of images, which are not restricted to anomalous structures, to provide a compact representation of entire volumes, and omit the need to prepare a dataset of normal patients. We believe that this makes our approach a valuable tool for hypothesis generation and biomarker identification, supporting a critical shift of mindset in medical image analysis. Namely, it expands the conventional biomarker evaluation strategy from supervised automation of expert annotation in known anomalies to an unrestricted unsupervised exploration of large-scale datasets.

Whenever image data are compressed to high level representations, topographic information of the represented biomarkers is reduced. In our second, global embedding, visualization of the features is not possible any longer because they do not contain any spatial information. Thus, it is challenging to interpret the individual biomarkers and their contributions to the variability of retinal morphology. These difficulties in interpreting the mechanisms of the deep learning model constitute the main limitation of our proposed algorithm. From a clinical perspective, it has always been desirable to clinically understand the steps taken by a model to reach a particular decision.(21) However, if doctors wish to augment their practice by artificial intelligence, these traditional paradigms may need to be revisited particularly as the hardware technology of image acquisition has long surpassed the feasibility of expert-based definition of features and assessment of imaging studies.

In this paper we introduced unsupervised deep learning to analyze high-resolution, three-dimensional retinal images without human-introduced bias. We presented novel auto-encoder based technology to capture the most relevant local structural biomarkers, including discovery of a new marker candidate. In a second embedding, we obtained a compact global description of the complex three-dimensional retinal scans, which nevertheless correlated better to visual acuity of patients than established artificial-intelligence based measurements. Once validated in additional, independent datasets, unsupervised machine learning and the resulting biomarkers may be employed in medical image analysis in retinal imaging and beyond.

4. Materials and Methods

4.1. Background and approach

In optical coherence tomography (OCT) an interferogram is obtained at a specific point of a sample, yielding an A-Scan containing one-dimensional information (along the z-axis) [9]. The A-scan data thus represent the condition of the retina at that specific position in the eye. By scanning the measurement beam across the sampling area, millions of A-scans are concatenated to form entire volume scans. It is this multi-step data acquisition which motivates the reasoning behind our proposed approach. Instead of trying to find an embedding for a volume in a single step, we construct two separate embeddings as depicted in Figure 2 that reflect the underlying process of OCT acquisition as well as the basic anatomy of the retina. In the first level, we learn a compact embedding of A-Scans and therefore of the local condition of the retina, using a fully connected auto-encoder. In the second level, a convolutional autoencoder is used to learn a global representation of whole OCT volumes based on the embedding obtained in the first level, resulting in a massive reduction of dimensionality.

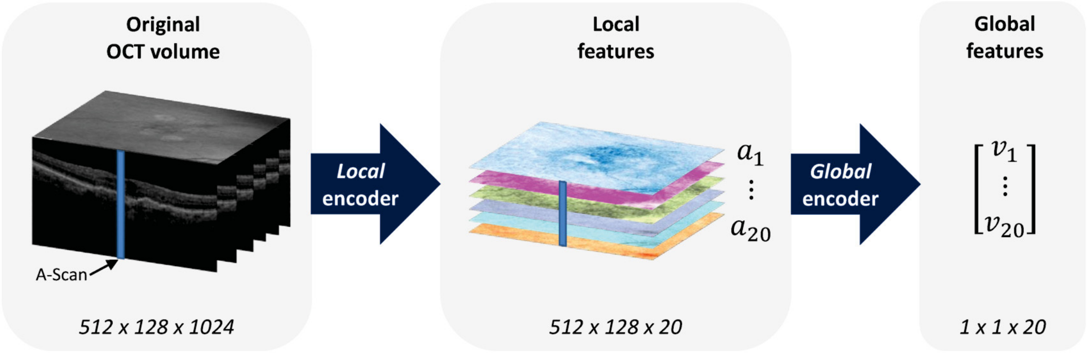


Figure 2: Flow-chart of the proposed two-level deep learning pipeline. In each step, an auto-encoder learns to encode the input data in a lower-dimensional embedding. First, the local encoder transforms each A-Scan into a 20-dimensional local representation, resulting in 20 2D feature maps. This local representation forms the input of the second stage, the global encoder. The global features provide a compact representation of an entire OCT volume.

4.2. Dataset

The experiments reported in this paper were conducted on a dataset consisting of 54,900 OCT volume scans of 1,094 patients enrolled in a randomized clinical trial [1]. The volumes were acquired using Cirrus OCT devices (Carl Zeiss Meditec, Dublin, CA, USA) and had a voxel dimensionality of 512 × 128 × 1024, covering a physical area of 6mm × 6mm × 2mm, with a voxel spacing of 11.7m × 46.9m × 2m. The dataset was randomly divided into a train (90%) and test set (10%) with 985 and 109 patients, respectively. There was no overlap of patients between those two sets. All study procedures were conducted in accordance with the tenets set forth in the Declaration of Helsinki and following Good Clinical Practice guidelines. All patients provided written informed consent before enrollment into the clinical trial. For the retrospective analysis of the image data, approval was obtained by the Ethics Committee at the Medical University of Vienna, Austria.

4.3. Data preprocessing

To reduce the large amount of speckle noise inherently present in OCT data, we use Bilateral Grids due to their fast runtime and easy implementation, on the individual B-Scans. We perform a single pass of filtering to reduce noise while retaining subtle details [2]. The position of the retina along the A-Scan is not fixed and depends on patient position during acquisition. To be invariant to this translation we compute a one-dimensional Fast Fourier Transform (FFT) of the A-Scan and discard the phase information by keeping only the magnitude of the complex FFT signal. Due to the resulting symmetry of the real-valued signal we only keep a vector of length 512 of the FFT amplitudes per 1024-long A-Scan.

4.4. Deep unsupervised learning of local features

Auto-encoders are trained without any labels and consist of two parts, the encoder and the decoder. During training, the input is encoded by the encoder into a low-dimensional embedding, and subsequently decoded by the decoder to reconstruct the original input. The underlying assumption is that the auto-encoder has to learn a meaningful compact high-level representation of the data to be able to perform accurate reconstruction. In Figure 3 and Figure 4, information within each auto-encoder always flows from the left to the right, with the embedding being the lowest-dimensional state in the middle of the stack. In the first stage of our framework (Figure 3), the A-Scan auto-encoder AE_1 is composed of three simple fully connected layers ([256/64/20] channels), with a weight matrix W_l, a bias vector b_l and an activation function σ:

$$y_l = \sigma(W_l x_l + b_l). \tag{1}$$

The sizes of the layer on both sides of the embedding are mirrored, and the weight matrices of two corresponding layers are tied: $W_l = W_l^T$. Throughout this work the activation function is set to be the exponential linear unit (ELU) [3], with $\alpha = 1$:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(exp(x) - 1) & \text{if } x \leq 0 \end{cases}, \tag{2}$$

The cost function used to drive the optimization in auto encoders measures the reconstruction error of the final output y given an input vector x:

$$C(x) = \sum (x - y)^T (x - y). \tag{3}$$

Using a randomly sampled subset of all the A-Scans available in the training set (1,600,000 A-Scans), a first auto-encoder is learnt in an end-to-end fashion as proposed in Zhou et al.28. The A-Scans within each volume are sampled from a Gaussian distribution, implying a higher chance for more centrally-located (clinically relevant) A-Scans to be part of the training subset. After training, only the encoder is used to map the A-Scans of all volumes into the embedding space, yielding the A- Scan features for each A-Scan. The individual A-Scans are processed independently from their position in and membership of any OCT volumes.

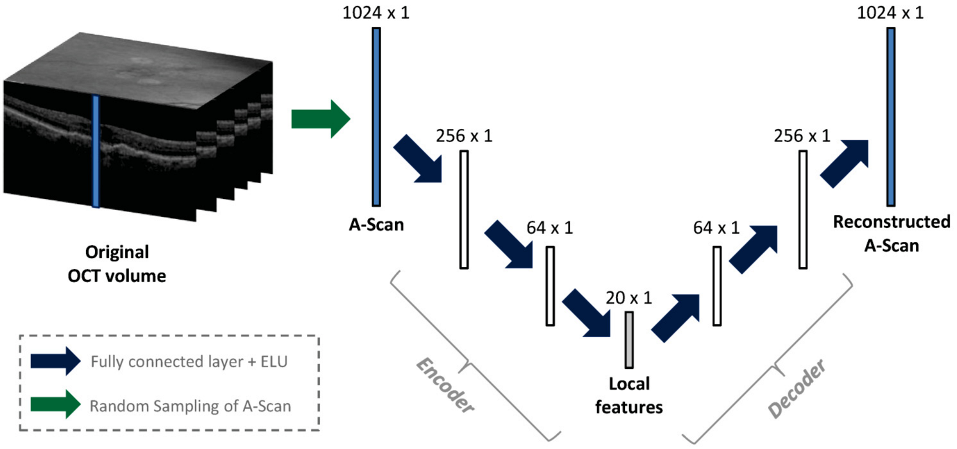


Figure 3: Illustration of the local auto-encoder architecture AE_1. Local features are learned using randomly sampled A-Scans from OCT volumes. During training, A-Scans are reconstructed from the compact representation (20 dimensions).

4.5. Deep unsupervised learning of global features

The A-Scan features of each volume are normalized feature-wise (zero mean, unit standard deviation) and concatenated according to their positions in the volume, yielding A-Scan feature volumes, reducing the volume size 50 times from 512 × 128 × 1024 to 512 × 128 × 20. Based on this compressed representation, in the second part of our framework, a deep convolutional auto-encoder is trained from all 49,505 training volumes.

This second auto-encoder AE_2 is composed of one linear down-sampling layer, followed by five convolutional ([64/64/126/256/512] channels) and three fully connected layers ([256/64/20] channels) on the encoder side, and a mirrored structure on the decoder side, as depicted in Figure 4. All layers are followed by the non-linear activation function ELU, and random-region dropout is applied to the input during training [11]. Applying the encoder of AE_2 on the A-Scan feature volumes yields a 20-dimensional global feature vector for each volume.

4.6. Statistical analysis evaluation using correlation

For evaluation of the learned features, the treatment nave baseline study eye OCT of each patient was used. Two different representations were selected as input for our quantitative evaluation:

- To enable a comparison with the same number of features, for each local A-Scan feature, the mean was calculated across the volume and used as feature representation of the OCT (20 dimensions).
- The global feature vector (20 dimensions).

We computed the Pearson correlation coefficient of the detected features with known markers, including markers of visual function (BCVA, LLVA) [4], retinal morphology as conventionally measured from OCT (retinal thickness, intra-retinal cystoid fluid, subretinal fluid, pigment epithelial detachment) and measures of disease activity obtained by fluorescein angiography (total area of lesions, total area of leakage) [18]. The correlation coefficients were computed utilizing the available information of all 1094

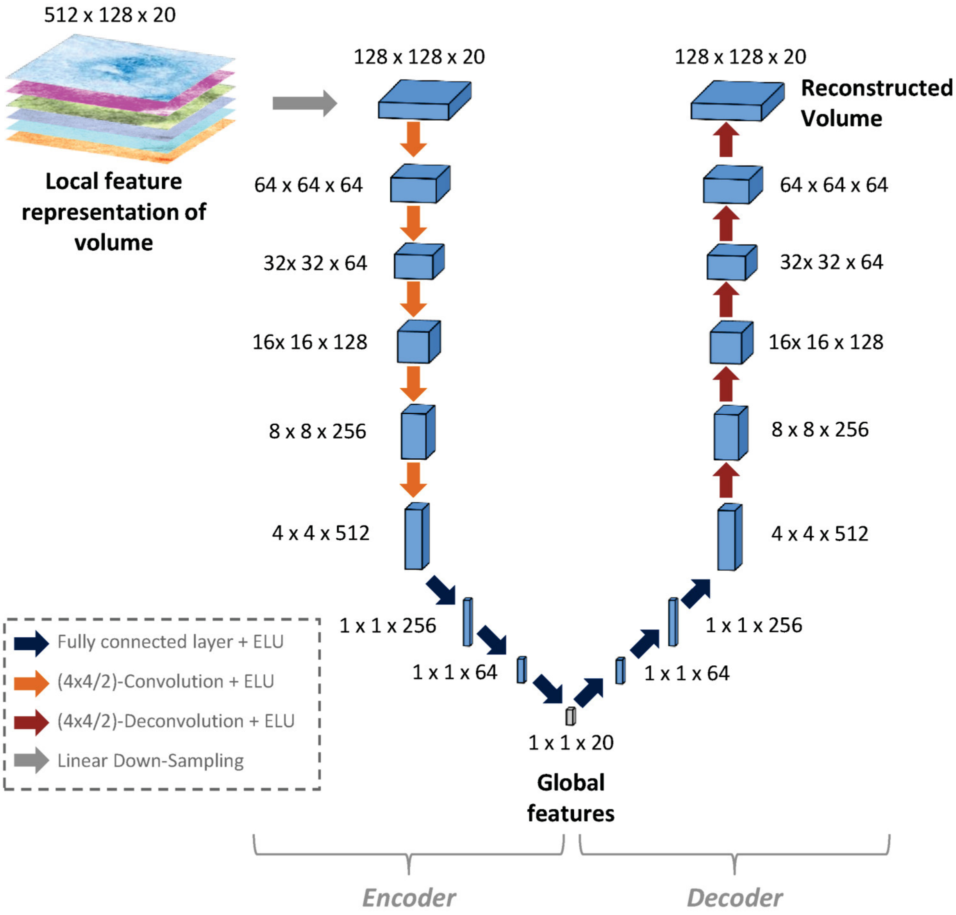


Figure 4: Architecture of the global auto-encoder AE_2. Encoding the local feature representation volume yields a compact global feature embedding, representing the whole OCT volume in only 20 dimensions.

patients. Additionally, we conducted hypothesis tests to evaluate if the correlation coefficients were significantly different from zero. Since this an explorative study, we did not perform correction for multiplicity testing in order not to increase the type II error (missing an effect that is present). Results are presented in Table 1 and Supplementary Table S1, where correlations with no significant difference from 0 are shown greyed out.

4.7. Evaluation using machine learning regression

For the evaluation using regression models, the treatment nave baseline study eye OCTs described above, where randomly divided into training and test set of 985 and 109 patients, respectively. We trained a multivariate linear regression model to predict the above-mentioned known markers. Elastic net regularization with 5-fold cross-validation was used to determine the optimal hyper-parameters ($\alpha = [0.0010.010.10.30.50.70.91]$). The performance of the final model was evaluated on the test set. For comparison, conventional OCT markers obtained by supervised deep learning, as described in Section 2.5 (21 dimensions), were used as variables to predict visual function (BCVA, LLVA), using the same settings for the linear regression model as described above. To test if the difference between the regression models (global features vs. conventional features) regarding the coefficient of determination was statistically significant, we performed a one-sided Wilcoxon signed-rank test.

4.8. Training details

For the training of the fully connected and the convolutional auto-encoder, we used Adam optimizer with standard parameters. For the former we used a learning rate of 0.0001, early stopping with a maximum of 500 epochs, a minibatch size of 64 and dropout at the input level with a rate of 0.5. For the latter we used a learning rate of 0.0001 for 10 epochs and 0.00001 for 2 epochs, a minibatch size of 8, random-region dropout-factor of 0.25 for the input and ordinary dropout in the first fully-connected layer of AE_2 with a factor of 0.5.

Acknowledgements

SMW, US-E, and GL are members of the medical imaging cluster (MIC) at the Medical University of Vienna.

Funding: The financial support by the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged.

Author Contributions: SMW, PS and GL designed the study, analyzed the data and wrote the article. PS, RD and GL developed the computer method. SMW, AS, AO and US-E contributed the dataset. All authors critically revised the article and agreed with submission for publication. SMW and US-E provided funding and take overall responsibility for the integrity of the data and analysis.

Competing interests: SMW serves as a consultant to Bayer and Novartis. AO is an employee of Genentech, Inc. USE serves as a consultant to Bhringer Ingelheim, Genentech, Novartis and Roche.

References

[1] Busbee, B. G., Ho, A. C., Brown, D. M., Heier, J. S., Suñer, I. J., Li, Z., Rubio, R. G., Lai, P., & Group, H. S. (2013). Twelve-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration. *Ophthalmology*, *120*, 1046–1056.

[2] Chen, J., Paris, S., & Durand, F. (2007). Real-time edge-aware image processing with the bilateral grid. In *ACM Transactions on Graphics (TOG)* (p. 103). ACM volume 26.

[3] Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, .

[4] Cocce, K. J., Stinnett, S. S., Luhmann, U. F., Vajzovic, L., Horne, A., Schuman, S. G., Toth, C. A., Cousins, S. W., & Lad, E. M. (2018). Visual function metrics in early and intermediate dry age-related macular degeneration for use as clinical trial endpoints. *American journal of ophthalmology*, *189*, 127–138.

[5] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., Van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A., & Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, *24*, 1342.

[6] Fujimoto, J., & Swanson, E. (2016). The development, commercialization, and impact of optical coherence tomography. *Investigative ophthalmology & visual science*, *57*, OCT1–OCT13.

[7] Hogg, R. E., Silva, R., Staurenghi, G., Murphy, G., Santos, A. R., Rosina, C., & Chakravarthy, U. (2014). Clinical characteristics of reticular pseudodrusen in the fellow eye of patients with unilateral neovascular age-related macular degeneration. *Ophthalmology*, *121*, 1748–1755.

[8] Holz, F. G., Sadda, S. R., Busbee, B., Chew, E. Y., Mitchell, P., Tufail, A., Brittain, C., Ferrara, D., Gray, S., Honigberg, L., Martin, J., Tong, B., Ehrlich, J. S., & Bressler, N. M. (2018). Efficacy and safety of lampalizumab for geographic atrophy due to age-related macular degeneration: Chroma and spectri phase 3 randomized clinical trials. *JAMA ophthalmology*, *136*, 666–677.

[9] Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A. et al. (1991). Optical coherence tomography. *science*, *254*, 1178–1181.

[10] Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., Karssemeijer, N., & Lillholm, M. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging*, *35*, 1322–1331.

[11] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536–2544).

[12] Puell, M. C., Barrio, A. R., Palomo-Alvarez, C., Gómez-Sanz, F. J., Clement-Corral, A., & Pérez-Carrasco, M. J. (2012). Impaired mesopic visual acuity in eyes with early age-related macular degeneration. *Investigative ophthalmology & visual science*, *53*, 7310–7314.

[13] Rohm, M., Tresp, V., Müller, M., Kern, C., Manakov, I., Weiss, M., Sim, D. A., Priglinger, S., Keane, P. A., & Kortuem, K. (2018). Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration. *Ophthalmology*, *125*, 1028–1036.

[14] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging* (pp. 146–157). Springer.

[15] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

[16] Schmidt-Erfurth, U., Bogunovic, H., Sadeghipour, A., Schlegl, T., Langs, G., Gerendas, B. S., Osborne, A., & Waldstein, S. M. (2018). Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmology Retina*, *2*, 24–30.

[17] Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., & Bogunović, H. (2018). Artificial intelligence in retina. *Progress in retinal and eye research*, .

[18] Schmidt-Erfurth, U., & Waldstein, S. M. (2016). A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Progress in Retinal and eye Research*, *50*, 1–24.

[19] Seeböck, P., Waldstein, S. M., Klimscha, S., Bogunovic, H., Schlegl, T., Gerendas, B. S., Donner, R., Schmidt-Erfurth, U., & Langs, G. (2018). Unsupervised identification of disease marker candidates in retinal oct imaging data. *IEEE Transactions on medical imaging*, .

[20] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*, 354.

[21] Wickström, K., & Moseley, J. (2017). Biomarkers and surrogate endpoints in drug development: a european regulatory view. *Investigative ophthalmology & visual science*, *58*, BIO27–BIO33.

[22] Willoughby, A. S., Ying, G.-s., Toth, C. A., Maguire, M. G., Burns, R. E., Grunwald, J. E., Daniel, E., & Jaffe, G. J. (2015). Subretinal hyperreflective material in the comparison of age-related macular degeneration treatments trials. *Ophthalmology*, *122*, 1846–1853.

List of Supplementary Materials

Table S1. Univariate Pearson correlation coefficients between global features and clinical target variables.

	Functional		OCT				Fluorescein Angiography	
	Best Corrected Visual Acuity	Low Luminance Visual Acuity	Retinal Thickness	Intra-Retinal Cystoid Fluid	Subretinal Fluid	Pigment Epithelial Detachment	Total Area of Lesion	Total Area of Leakage
a1	-0.21	-0.19	0.30	0.07	0.25	0.15	0.17	0.17
a2	-0.20	-0.17	0.36	0.20	0.32	0.15	0.11	0.13
a3	-0.10	-0.08	0.39	0.18	0.27	0.10	0.06	0.08
a4	0.30	0.36	-0.32	-0.07	-0.22	-0.24	**-0.32**	**-0.33**
a5	**-0.31**	**-0.40**	0.30	0.20	0.33	0.19	0.25	0.31
a6	-0.10	-0.03	0.31	0.07	0.19	0.12	0.07	0.05
a7	-0.08	-0.20	-0.05	-0.00	0.17	0.17	0.20	0.19
a8	-0.08	-0.04	0.29	0.14	0.24	0.11	0.06	0.06
a9	-0.17	-0.21	0.27	0.05	0.15	0.09	0.10	0.17
a10	-0.29	-0.27	0.54	0.20	0.41	0.24	0.22	0.24
a11	-0.05	-0.03	0.03	-0.02	0.11	0.07	0.06	0.04
a12	-0.08	0.00	0.49	0.18	0.26	0.10	-0.01	0.02
a13	-0.22	-0.26	0.41	0.27	0.39	0.22	0.18	0.19
a14	0.05	0.03	-0.31	-0.04	-0.04	-0.09	-0.10	-0.10
a15	0.03	0.02	0.01	-0.04	0.08	-0.01	0.08	0.08
a16	-0.02	-0.01	0.08	-0.04	0.16	0.10	0.09	0.05
a17	-0.28	-0.23	**0.73**	**0.37**	**0.45**	**0.32**	0.23	0.26
a18	-0.23	-0.26	0.18	0.00	0.24	0.17	0.22	0.22
a19	-0.19	-0.16	0.28	0.10	0.28	0.15	0.13	0.12
a20	-0.22	-0.19	0.39	0.18	0.33	0.21	0.16	0.17

Supplementary Table S1: Univariate Pearson correlation coefficients between the global features (v1-v20) and functional variables as well as measures of disease activity by OCT and fluorescein angiography. Green colour indicates a positive, and blue colour a negative correlation. Correlations with no significant difference from 0 are shown greyed out.

CHAPTER 8

Discussion and Conclusion

"The ultimate ignorance is the rejection of something you know nothing about, yet refuse to investigate."

– Wayne Dyer

In this final Chapter the presented work is summarized, including a discussion of the methodology, findings, limitations and potential use in a broader context. First, we provide an overview about the status quo as well as the potential of machine learning in general. Thereafter, the contributions of this thesis are summarized and discussed. Finally, a review of potential limitations of the presented methods and a discussion of possible future lines of research concludes this Chapter.

8.1 The potential of machine learning

The success of deep learning is based on various aspects such as the availability of ”big data”, the computational capacity of modern hardware and the methodological development in the field of of machine learning. Huge amounts of data do not provide knowledge on their own, but rather need to be processed by appropriate machine learning algorithms in order to exploit potential. In particular, deep learning is a powerful tool to transform data into knowledge, since it can learn extremely complex functions directly from data, in an end-to-end manner (Litjens et al., 2017), making it one of the critical tools to advance and practice medicine (Obermeyer and Emanuel, 2016).

In medicine, various disciplines are affected by machine learning algorithms. In particular, the potential to perform automated screening, classification or prediction of deep learning based approaches on large-scale biomedical image data has been demonstrated. For instance, in Poplin et al. (2018) the risk of cardiovascular diseases has been predicted from retinal fundus photographs. Hospital admissions in emergency departments have been robustly predicted from patient history and triage information in Hong et al. (2018). The automatic classification of echo-cardiograms is another example regarding the application of deep learning in medicine (Madani et al., 2017). Moreover, machine learning has a remarkable potential of improving healthcare in settings where medical or financial resources are limited, e.g. in remote areas or third-world countries (Koch, 2018). In this context, automatic screening of common diseases would be one promising area of application, including automated classification of skin cancer (Esteva et al., 2017) or detection of diabetic retinopathy from fundus photographs (Carson Lam et al., 2018, Gulshan et al., 2016).

Machine learning is expected to (1) increase diagnostic accuracy, (2) improve prognosis and (3) displace much of the physicians work with respect to interpreting digitized images by improvements in medical image analysis (Obermeyer and Emanuel, 2016). Doctors have to handle enormous amounts of data in clinical medicine, with raising complexity of this task due to increasing dimensionality of data (e.g. images) and new medical technologies. At this point, machine learning will become an important tool for clinicians to understand their patients (Obermeyer and Emanuel, 2016). However, the task at this stage is not to replace but rather to support clinicians, which leads to the challenge of integrating these methods into the clinical workflow and existing structures (Koch, 2018).

In this context, machine learning methods that are based on supervised learning allow to automatize the task of detecting specific lesions or to predict pre-defined outcome variables. While supervised deep learning methods can achieve human-level performance, they require a large amount of labeled data, which are costly or sometimes unfeasible to obtain. Additionally, they are restricted to a-priori known disease categories or lesions which limits their exploratory power. At the same time, the discovery of effective new biomarkers is important to improve diagnosis, treatment and management of patients in general. Here, the development of appropriate machine learning approaches that allow for exploration of medical imaging data without needing manual labels is crucial. These methods can help to improve individual patient care by discovering new biomarkers, constituting a shift towards hypothesis-generation centered strategies.

8.2 Contribution

In this thesis we have presented several contributions for biomarker discovery in medical images, focusing on investigating patient populations suffering from age-related macular degeneration (AMD). Although the application of the proposed approaches is not limited to this disease, AMD is highly relevant due to the fact that it is the most common cause for severe vision loss and blindness in industrialized countries. At the same time, the existing knowledge gaps with respect to the pathogenesis of retinal diseases and the absence of effective treatments for certain disease patterns necessitates the discovery of informative biomarkers.

Moreover, biomarkers are needed to improve early diagnosis, optimize existing treatment strategies and enhance the management of patients. As OCT provides non-invasive high-resolution imaging of the retinal morphology, the investigation of this image data constitutes a promising field of research to identify relevant new markers.

In the main part of the thesis, the potential of learning retinal features from large-scale image data without manual labels was explored. New approaches for detecting anomalies and categorizing new marker candidates have been proposed. We validated the discriminative power of newly identified marker candidates for the differentiation of separate disease stages, showed the ability of Bayesian deep learning for anomaly detection and demonstrated the improved predictive capability of learned features compared to conventional known disease markers with respect to functional outcome measures. We elaborated three strategies to identify new biomarker candidates in medical image data, which are summarized in the following.

Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data In Chapter 5, we presented an unsupervised approach to identify and categorize disease marker candidates in retinal OCTs on pixel-level, based on anomaly detection and subsequent clustering. A multi-scale autoencoder and a One-class SVM are trained on a set of healthy training images to learn the distribution of normal appearance, allowing the detection of anomalies on new images. This can be viewed as a semantic segmentation approach, where all pixels not in the healthy appearance distribution are classified as abnormal. A subsequent clustering step yields a categorization of the anomalies into sub-groups, serving as disease marker candidates. The evaluation results revealed comparable or superior performance to alternative unsupervised feature learning techniques regarding the segmentation of anomalous regions in OCT scans. We demonstrated that the model identified stable categories that were replicable across datasets. At the same time, the results showed that disease processes were captured by the anomalous categories, suggesting that valuable discriminative information is encoded in the found marker candidates, indicating their link to disease.
Clinical prospects: This work has introduced a novel approach for unsupervised identification of novel marker candidates and can serve as a hypothesis-generation strategy (cf. Section 4). Since known biomarkers such as retinal thickness or macular fluid neither explain the full spectrum of the disease nor the individual level of vision loss, this work is as a step towards tools for a better understanding of retinal diseases, and therefore for enhanced patient care. A set of healthy training cases is the only requirement to enable the training of the proposed unsupervised anomaly detection method. This means that the approach is expected to be broadly applicable to any type of lesion, disease or biomedical imaging data, since it neither relies on lesion-specific annotations nor on pixel-wise or volume-wise classification labels.

Epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT A novel anomaly detection approach based on epistemic uncertainty estimations has been presented in Chapter 6. A deep network is trained on a cohort of normal subjects to segment the healthy anatomy, i.e. the layers of the retina. The weak labels of the retinal layers used for training the Bayesian deep learning model were generated using an automated method, taking advantage of the fact that traditional approaches are expected to work accurately in a well-defined environment, i.e. in a set of healthy cases. In this way, information about healthy anatomical appearance is injected into the model without the need for manual annotations. The usage of weak labels allows to produce more training data and thereby to capture more healthy appearance variability. During test time, deviations from normal anatomical variability are detected using epistemic uncertainty estimates, which are expected to correlate with these anomalies not observed during training. This is supported by the results, showing a high correlation between the area of anomaly annotations and uncertainty estimates. To obtain smooth binary segmentation maps of anomalies, the resulting uncertainty maps are processed with a newly introduced majority-ray-casting technique. A quantitative evaluation revealed that the proposed method clearly outperformed the unsupervised baseline approach with respect to the segmentation of anomalous regions in retinal OCT scans. The importance of target labels that describe the healthy anatomy in an informative way is highlighted by results of ablation experiments, which demonstrated that using less informative target labels (i.e. binary segmentation of the whole retina instead of its individual layers) resulted in a decreased performance for anomaly detection.

Clinical prospects: In general, using weak labels to inject knowledge about healthy anatomical variation into the model, combined with the power of Bayesian deep learning, offers an effective tool for anomaly detection in biomedical imaging data. Since the detected anomalies are either new biomarker candidates or known disease markers, the possible applications are two-fold. In the first case, the marker candidates can be analyzed in a clinical post-hoc analysis, which means that the anomaly detection model plays an important role regarding hypothesis-generation. In the second case, the anomaly detection model can be utilized more directly in a clinical setting to identify abnormal images, e.g. for screening. The received abnormality-scores for biomedical images could be used to select samples for a subsequent revision by clinicians, referring cases with anomalies for further analysis. In both cases, we expect the anomaly detection model to generalize well to all kind of diseases that alter the anatomical appearance, not restricted to a specific pathology.

A paradigm shift in retinal biomarker identification by unsupervised deep learning In Chapter 7, the focus is shifted from anomaly detection to unsupervised feature learning in OCTs. This constitutes a different strategy in terms of biomarker discovery. In contrast to anomaly detection, the target is not to learn the appearance of normal cases and detect and analyze all deviations from normal, but instead learn the most prominent features from a large-scale population of diseased patients without labels, and analyze disease specific image characteristics in the space of these features. This not only omits the need for manual labels of specific lesions or target classes, but also eliminates the requirement of collecting a set of healthy cases, which can be challenging in some clinical settings. Moreover, the former allows to explore the data in an unbiased way, since the learned features are not adjusted towards a specific prediction task. In two separate stages, an auto-encoder based approach is trained on a population of AMD patients to learn a low-dimensional representation of the local morphology of the retina, as well as a global embedding of whole OCT volumes. This explicit division of the architecture into multiple stages allows to conduct analyses both on a local A-scan and on a global volume level. A quantitative and qualitative evaluation revealed that conventionally used biomarkers (e.g. retinal thickness) were captured by the local features. This demonstrates that our method is capable of identifying clinically meaningful morphological information. In addition, the new local embedding also learned features that had not been considered yet in clinical practice (i.e. features that form new biomarker candidates), highlighting the potential of the proposed approach to generate new findings. Moreover, the results showed that the learned global features outperformed conventional OCT markers with respect to the task of predicting visual function in our AMD patient population.

Clinical prospects: Due to the high-dimensional property of three-dimensional OCT volumes, the manual investigation of this data is an extremely complex task. At the same time, analyzing large amounts of data and generating labels is costly or sometimes even unfeasible. In contrast, the proposed method allows to automatically learn features from large-scale imaging data without the need of manual annotations at multiple levels of granularity (local and global features). For instance, one local feature that showed the highest correlation with functional target variables (BCVA, LLVA), could be qualitatively linked with subretinal hyperreflective lesions, which are currently not used as endpoint in clincial trials. This demonstrates the potential of the proposed method to learn new marker candidates, exploring the population of diseased patients in an unbiased way, without labels. The global features on the other hand allow to compress the complex morphological information of a 3D retinal image volume into an extremely compact representation. This offers a new perspective on how to explore and make high-dimensional medical imaging data manageable in the near future.

8.3 Future lines of research

Several lines of future research can be derived from the proposed methods, discussed in the following.

Holistic versus isolated anomaly detection approaches for biomarker discovery The proposed unsupervised anomaly detection approaches allowed to identify categories of anomalies (i.e. disease marker candidates) that encode discriminative information regarding disease. However, the objective of identifying categories restrained the approach to a localized representation that enabled clustering on a local level. This in turn led to a limitation with respect to the anomaly detection performance, which left room for improvement. We tackled this issue in our second work (Chapter 6) where we developed an approach focusing solely on the task of anomaly detection. Recently, also other work (e.g. Schlegl et al. (2017)) has been proposed to tackle the problem of anomaly detection as an isolated task. In general, the identification of disease marker candidates (as done in Chapter 5) constitutes a first step in the process of biomarker discovery. Subsequently, a precise description of characteristics of those candidates must be created. This is necessary to enable a transformation from candidates to effective markers applicable in clinical practice in a consecutive post-hoc analysis. In this context, an open research question that remains is how to optimize the process of biomarker discovery using anomaly detection. On one hand, considering anomaly detection as an isolated task may limit the applicability of the developed method for the necessary subsequent steps. Furthermore, tackling the individual steps separated from each other may ignore the existing potential, lying in the fusion of multiple tasks. On the other hand, aiming for a holistic approach of the biomarker discovery pipeline may restrain the individual components from a methodological perspective. Since the anomaly detection performance was substantially increased by the work presented in Chapter 6, future work could either aim at integrating an anomaly categorization step directly into the model, or developing a separate method that uses the outcome of the proposed approach as input.

Definition of normal training set One advantage of the presented anomaly detection methods is that they require only a set of healthy cases for training. This omits the need for costly large-scale manual annotations of a-priori defined lesions and increases the generalizability of the model to all kinds of diseases, since all deviations from normal are detected by definition. Nevertheless, the performance of anomaly detection approaches can be limited by the quality of the available healthy training samples. In connection with biomarker discovery, it is important to ensure that no anomalies are present in the training set, since they would be incorporated as normal appearance rather than being detected as anomalies. Moreover, the training set should cover the entire spectrum of healthy appearance, as normal cases could otherwise be erroneously detected as anomalous. Furthermore, in some clinical scenarios it can be challenging to obtain a set of healthy cases, since medical image acquisition is usually motivated by a certain clinical condition.

Quantitative evaluation While the quantitative evaluation of anomaly segmentation results provides an objective measurement of performance, the explanatory power is limited.

This is caused by the fact that manually annotating anomalies is a difficult task: transitions between healthy and diseased scans are continuous, hard to define, often unclear and exposed to subjective interpretation. Ensuring consistent ground-truth labels is therefore nearly impossible, especially in the border regions of anomalous areas. For instance, a low precision value may be the result of either normal appearance that has not been captured by the anomaly detection model, or emerge from anomalies that have not yet been categorized as such in the manual annotations and are potential new marker candidates. A careful evaluation is therefore crucial. One way to tackle this limitation has been shown in Chapter 6, where a lesion-wise evaluation score has been used to provide a more comprehensive view on the model performance. This kind of evaluation allowed to link quantitative pixel-based evaluation metrics with lesion-level detection capabilities. More generally, the assessment of inter- and intra-reader variability constitutes an important step towards understanding the complexity of the given task. Examining the variability of readers would therefore be another interesting line of future research.

Unsupervised Representation Learning As mentioned above, the learned compact disease specific representation of a medical image allows to analyze the amount of information contained in the images. We demonstrate that the obtained features correlate better with clinically relevant measures of retinal function and disease activity than established biomarkers. However, since the features are learned without labels, they are quite general in the sense that they capture only the most prominent characteristics of the dataset, i.e. the patient population. This means that the model is prone to miss some features which are not so prominent but are still important for a specific task at the same time. This is related to one of the main challenges in representation learning, namely the difficulty of establishing a clear target to train the machine learning model (cf. Section 3.1.2). One possible solution would be to fine-tune the representation, training the network with a small subset of labeled cases in a consecutive step. Another interesting research line would be to combine anomaly detection and representation learning: After detecting all anomalies in a first step, representation learning is then conducted using only the abnormal regions. This could facilitate the process of discovering the underlying structure and characteristic of diseases.

Post-hoc analysis In general the analysis, evaluation and interpretation of marker candidates remains challenging and complex. Since they do not correspond to known categories, there exists no ground-truth for their direct evaluation. Instead, we conducted qualitative evaluation by experts as well as correlation, regression and classification experiments to verify and investigate the nature of learned features, alias marker candidates. While qualitative evaluation allows an in-depth exploration of results but is time-consuming, the other evaluation approaches can suffer from noise in the target labels (cf. Section 2.1.2) or may be biased by the choice of the used target variable itself. In this context it seems all the more meaningful to ”decouple” the process of identifying new marker candidates from the evaluation as much as possible, as this minimizes the bias towards established medical hypotheses. In this thesis, we propose different approaches that allow to explore data and identify new biomarker candidates, but at the same time do not rely on predefined hypotheses, lesion-descriptions or outcome variables. However, more work is needed in order to enhance the process of biomarker discovery supported by machine learning methodology.

8.4 Conclusion

This dissertation proposes new machine learning methods that tackle the problem of identifying disease marker candidates in different ways: (1) a novel anomaly detection approach that identified and categorized anomalies in an unsupervised way, (2) an anomaly detection method that is based on Bayesian deep learning, and (3) a completely unsupervised feature learning approach that captures phenotype characteristics of the patient population. Even though the methods were evaluated on a specific disease (AMD), we expect them to generalize well to other retinal diseases. We believe that the published work represents a step towards improving individual patient care by proposing new ways of identifying new imaging biomarkers, using the abilities of machine learning.

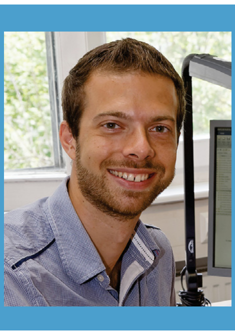
Appendix

Curriculum vitae

DI Philipp **Seeböck**

Westbahnstrasse 1 | 1070 Wien | Austria

p.seeboeck@gmx.net



Education

2015-present **DOCTOR OF PHILOSOPHY** *Medical University of Vienna, Austria*
Thesis Title: Discovery of Biomarker Candidates in Retinal OCT Images using Deep Learning. Supervisor: Dr. Georg Langs.

2012-2015 **DIPLOM-INGENIEUR** *in Medical Informatics, Vienna University of Technology, Austria.*
Thesis title: Deep Learning In Medical Image Analysis.
Passed with distinction.

2009-2012 **BACHELOR OF SCIENCE** *in Medical Informatics, Vienna University of Technology, Austria.*
Passed with distinction.

Research experience

2015-present **RESEARCH ASSOCIATE** *Doctoral Level, Computational Imaging Research Lab (CIR), Department of Biomedical Imaging and Image-guided Therapy, and Christian Doppler Laboratory for Ophthalmic Image Analysis (OPTIMA), Department of Ophthalmology and Optometry, Medical University Vienna, Austria.*

- Machine learning, Deep learning.
- Unsupervised Learning.
- Medical Image Analysis.
- Biomarker discovery.
- Application of Machine Learning methods on biomedical data.

2013-2015 **RESEARCH ASSISTANT** *Master's Level, Computational Imaging Research Lab (CIR), Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Austria.*

- Medical Image Analysis.
- Deep Learning.
- Unsupervised representation learning on thorax CT scans.

Awards

2019 **ARVO International Travel Grant**

2017 **Best Paper Award** at the Computer Vision Winter Workshop

Professional experience

2010-2015 **PROMOTER** *Brandfan Werbe GmbH, Vienna, Austria.*
Worked for the advertising company at trade fairs, fashion shows, etc.
Duties included sales promotion, stall support, customer service and taking care of contests

2008-2009 **COMMUNITY SERVICE (AS ALTERNATIVE TO MILITARY SERVICE)** *Wiener Hilfswerk, Vienna, Austria.*

List of publications

Peer-reviewed journal publications

1. **Seeböck, P.**, Waldstein, S.M., Klimscha. S., Bogunovic, H., Schlegl, T., Gerendas, B.S., Donner, R., Schmidt-Erfurth, U., Langs, G. (2019). *Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data.* IEEE Transactions on Medical Imaging, 38(84), 1037-1047.
2. Schlegl, T., **Seeböck, P.**, Waldstein, S.M., Langs, G., Schmidt-Erfurth, U. (2019) *f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks.* Medical Image Analysis.

Peer-reviewed conference publications

1. **Seeböck, P.**, Waldstein, S., Klimscha, S., Gerendas, B. S., Donner, R., Schlegl, T., Schmidt-Erfurth, U., Langs, G. (2016). *Identifying and categorizing anomalies in retinal imaging data.* NIPS 2016 MLHC workshop. preprint arXiv:1612.00686.
2. Schlegl, T., **Seeböck, P.**, Waldstein, S. M., Schmidt-Erfurth, U., Langs, G. (2017). *Unsupervised anomaly detection with generative adversarial networks to guide marker discovery.* In: Proceedings of Information Processing in Medical Imaging 2017. Preprint arXiv:1703.05921.
3. **Seeböck, P.**, Romo-Bucheli, D., Waldstein, S. M., Bogunović, H., Orlando, J.I., Gerendas, B. S., Langs, G., Schmidt-Erfurth, U. (forthcoming). *Using CycleGANs for Effectively Reducing Image Variability Across OCT Devices and Improving Retinal Fluid Segmentation.*
4. Orlando, J.I., **Seeböck, P.**, Bogunović, H., Klimscha, S., Grechenig C., Waldstein, S.M., Gerendas, B.S., Schmidt-Erfurth, U. (forthcoming). *U2-Net: A Bayesian U-Net Model with Epistemic Uncertainty Feedback for Photoreceptor Layer Segmentation in Pathological OCT Scans.*
5. **Seeböck, P.**, Donner, R., Schlegl, T., Langs, G. (2017). *Unsupervised learning for image category detection.* Computer Vision Winter Workshop.

Other conference publications, abstracts & posters

1. **Seeböck, P.**, Waldstein, S.M., Donner, R., Sadeghipour, A., Langs, G., Gerendas, B.S., Osborne, A., Schmidt-Erfurth, U. (2018). *Unsupervised deep learning to identify markers in optical coherence tomography.* Investigative Ophthalmology & Visual Science July, 59(9), 1736-1736.
2. **Seeböck, P.**, Waldstein, S.M., Donner, R., Gerendas, B.S., Sadeghipour, A., Osborne, A., Schmidt-Erfurth, U., Langs, G., (2017). *Defining disease endophenotypes in neovascular AMD by unsupervised machine learning of large-scale OCT data.* Investigative Ophthalmology & Visual Science, 58(8), 56-56.
3. Schlegl, T., Bogunovic, H., Klimscha, S., **Seeböck, P.**, Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., Langs, G., Schmidt-Erfurth, U. (2018). *Fully automated segmentation of hyperreflective foci in optical coherence tomography images.* arXiv preprint arXiv:1805.03278.

Bibliography

Abràmoff, M. D., Garvin, M. K., Sonka, M., 2010. Retinal imaging and image analysis. IEEE reviews in biomedical engineering 3, 169–208.

Achterberg, H. C., van der Lijn, F., den Heijer, T., Vernooij, M. W., Ikram, M. A., Niessen, W. J., de Bruijne, M., 2014. Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Human brain mapping 35 (5), 2359–2371.

Acton, J. H., Greenstein, V. C., 2013. Fundus-driven perimetry (microperimetry) compared to conventional static automated perimetry: similarities, differences, and clinical applications. Canadian Journal of Ophthalmology 48 (5), 358–363.

Al-Zamil, W. M., Yassin, S. A., 2017. Recent developments in age-related macular degeneration: a review. Clinical interventions in aging 12, 1313.

Ando, S., 2007. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, pp. 13–22.

Arden, G., Jacobson, J., 1978. A simple grating test for contrast sensitivity: preliminary results indicate value in screening for glaucoma. Investigative ophthalmology & visual science 17 (1), 23–32.

Arditi, A., Cagenello, R., 1993. On the statistical reliability of letter-chart visual acuity measurements. Investigative ophthalmology & visual science 34 (1), 120–129.

Asman, A. J., Landman, B. A., 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE transactions on medical imaging 30 (10), 1779–1794.

Awh, C. C., Hawken, S., Zanke, B. W., 2015. Treatment response to antioxidants and zinc based on cfh and arms2 genetic risk allele number in the age-related eye disease study. Ophthalmology 122 (1), 162–169.

Awh, C. C., Zanke, B., Kustra, R., 2017. Progression from no amd to intermediate amd as influenced by antioxidant treatment and genetic risk: an analysis of data from the age-related eye disease study cataract trial. Journal of VitreoRetinal Diseases 1 (1), 45–51.

Bailey, I. L., Lovie, J. E., 1976. New design principles for visual acuity letter charts. American journal of optometry and physiological optics 53 (11), 740–745.

Balasubramanian, S., Uji, A., Lei, J., Velaga, S., Nittala, M., Sadda, S., 2018. Interdevice comparison of retinal sensitivity assessments in a healthy population: the centervue maia and the nidek mp-3 microperimeters. British Journal of Ophthalmology 102 (1), 109–113.

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning. pp. 37–49.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2018. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: International MICCAI Brainlesion Workshop. Springer, pp. 161–169.

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade. Springer, pp. 437–478.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35 (8), 1798–1828.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. In: Advances in neural information processing systems. pp. 153–160.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5 (2), 157–166.

Bengio, Y., et al., 2009. Learning deep architectures for ai. Foundations and trends® in Machine Learning 2 (1), 1–127.

Bhutto, I. A., McLeod, D. S., Hasegawa, T., Kim, S. Y., Merges, C., Tong, P., Lutty, G. A., 2006. Pigment epithelium-derived factor (pedf) and vascular endothelial growth factor (vegf) in aged human choroid and eyes with age-related macular degeneration. Experimental eye research 82 (1), 99–110.

Bishop, C. M., et al., 1995. Neural networks for pattern recognition. Oxford university press.

Bogunović, H., Montuoro, A., Baratsits, M., Karantonis, M. G., Waldstein, S. M., Schlanitz, F., Schmidt-Erfurth, U., 2017. Machine learning of the progression of intermediate age-related macular degeneration based on OCT imaging. Investigative ophthalmology & visual science 58 (6), BIO141–BIO150.

Boureau, Y.-L., Bach, F., LeCun, Y., Ponce, J., 2010a. Learning mid-level features for recognition.

Boureau, Y.-L., Ponce, J., LeCun, Y., 2010b. A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 111–118.

Boyer, D. S., Schmidt-Erfurth, U., van Lookeren Campagne, M., Henry, E. C., Brittain, C., 2017. The pathophysiology of geographic atrophy secondary to age-related macular degeneration and the complement pathway as a therapeutic target. Retina (Philadelphia, Pa.) 37 (5), 819.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Bruna, J., Szlam, A., LeCun, Y., 2013. Signal recovery from pooling representations. arXiv preprint arXiv:1311.4025.

Bryll, R., Gutierrez-Osuna, R., Quek, F., 2003. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern recognition 36 (6), 1291–1302.

BÜhren, J., Terzi, E., Bach, M., Wesemann, W., Kohnen, T., 2006. Measuring contrast sensitivity under different lighting conditions: comparison of three tests. Optometry and vision science 83 (5), 290–298.

Carkeet, A., 2001. Modeling logmar visual acuity scores: effects of termination rules and alternative forced-choice options. Optometry and vision science 78 (7), 529–538.

Carson Lam, D. Y., Guo, M., Lindsey, T., 2018. Automated detection of diabetic retinopathy using deep learning. AMIA Summits on Translational Science Proceedings 2017, 147.

Chalapathy, R., Menon, A. K., Chawla, S., 2018. Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR) 41 (3), 15.

Channa, R., Smith, M., Campochiaro, P. A., 2011. Treatment of macular edema due to retinal vein occlusions. Clinical ophthalmology (Auckland, NZ) 5, 705.

Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-supervised learning.

Chappelow, A. V., Tan, K., Waheed, N. K., Kaiser, P. K., 2012. Panretinal photocoagulation for proliferative diabetic retinopathy: pattern scan laser versus argon laser. American journal of ophthalmology 153 (1), 137–142.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., Chen, C.-M., 2016. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Scientific reports 6, 24454.

Chew, E. Y., Clemons, T. E., Bressler, S. B., Elman, M. J., Danis, R. P., Domalpally, A., Heier, J. S., Kim, J. E., Garfinkel, R., Group, A.-H. S. R., et al., 2014. Randomized trial of a home monitoring system for early detection of choroidal neovascularization home monitoring of the eye (home) study. Ophthalmology 121 (2), 535–544.

Cho, S.-H., Jeon, J., Kim, S. I., 2012. Personalized medicine in breast cancer: a systematic review. Journal of breast cancer 15 (3), 265–272.

Chow, C., 1970. On optimum recognition error and reject tradeoff. IEEE Transactions on information theory 16 (1), 41–46.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Ciresan, D., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems. pp. 2843–2851.

Cireşan, D., Meier, U., Masci, J., Schmidhuber, J., 2011. A committee of neural networks for traffic sign classification. In: Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, pp. 1918–1921.

Cireşan, D., Meier, U., Masci, J., Schmidhuber, J., 2012. Multi-column deep neural network for traffic sign classification. Neural networks 32, 333–338.

Cireşan, D. C., Meier, U., Gambardella, L. M., Schmidhuber, J., 2010. Deep, big, simple neural nets for handwritten digit recognition. Neural computation 22 (12), 3207–3220.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). arXiv preprint arXiv:1511.07289.

Clifton, L., Clifton, D. A., Watkinson, P. J., Tarassenko, L., 2011. Identification of patient deterioration in vital-sign data using one-class support vector machines. In: 2011 federated conference on computer science and information systems (FedCSIS). IEEE, pp. 125–131.

Coates, A. P., 2012. Demystifying unsupervised feature learning. Ph.D. thesis, Stanford University.

Cole, E. D., Ferrara, D., Novais, E. A., Louzada, R. N., Waheed, N. K., 2016. Clinical trial endpoints for optical coherence tomography angiography in neovascular age-related macular degeneration. Retina 36, S83–S92.

Colijn, J. M., Buitendijk, G. H., Prokofyeva, E., Alves, D., Cachulo, M. L., Khawaja, A. P., Cougnard-Gregoire, A., Merle, B. M., Korb, C., Erke, M. G., et al., 2017. Prevalence of age-related macular degeneration in europe: the past and the future. Ophthalmology 124 (12), 1753–1763.

Collin, H. B., 2008. Is bcva an invention of ophthalmology? Clinical and Experimental Optometry 91 (5), 425–426.

Commons, W., 2015. Logmarvar: Wikimedia commons. [Online; accessed 9-February-2019].

Commons, W., 2018a. Retina: Wikimedia commons. [Online; accessed 28-January-2019].

Commons, W., 2018b. Schematic diagram of the human eye: Wikimedia commons. [Online; accessed 28-January-2019].

Council, N. R., et al., 2002. Visual impairments: determining eligibility for social security benefits. National Academies Press.

Curcio, C. A., Sloan, K. R., Kalina, R. E., Hendrickson, A. E., 1990. Human photoreceptor topography. Journal of comparative neurology 292 (4), 497–523.

Del Giorno, A., Bagnell, J. A., Hebert, M., 2016. A discriminative framework for anomaly detection in large videos. In: European Conference on Computer Vision. Springer, pp. 334–349.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? does it matter? Structural Safety 31 (2), 105–112.

Dice, L. R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dietterich, T. G., Lathrop, R. H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence 89 (1-2), 31–71.

Ding, J., Wong, T. Y., 2012. Current epidemiology of diabetic retinopathy and diabetic macular edema. Current diabetes reports 12 (4), 346–354.

Dobson, A. J., Barnett, A. G., 1990. An introduction to generalized linear models, 21–236.

Doersch, C., Gupta, A., Efros, A. A., 2015. Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1422–1430.

Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics 31 (4-5), 198–211.

Draper, N. R., Smith, H., 2014. Applied regression analysis. Vol. 326. John Wiley & Sons.

Drexler, W., Fujimoto, J. G., 2008. Optical coherence tomography: technology and applications. Springer Science & Business Media.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12 (Jul), 2121–2159.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification.

Duncan, J. S., Ayache, N., 2000. Medical image analysis: Progress over two decades and the challenges ahead. IEEE transactions on pattern analysis and machine intelligence 22 (1), 85–106.

Elliott, D. B., 2016. The good (logmar), the bad (snellen) and the ugly (bcva, number of letters read) of visual acuity measurement. Ophthalmic and Physiological Optics 36 (4), 355–358.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition.

Ergen, T., Mirza, A. H., Kozat, S. S., 2017. Unsupervised and semi-supervised anomaly detection with lstm neural networks. arXiv preprint arXiv:1710.09207.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research 11 (Feb), 625–660.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.

European Society of Radiology, 2015. Medical imaging in personalised medicine: a white paper of the research committee of the european society of radiology (esr). Insights into imaging 6 (2), 141–155.

Fahy, J., Glynn, D., Hutchinson, M., 1989. Contrast sensitivity in multiple sclerosis measured by cambridge low contrast gratings: a useful clinical test? Journal of Neurology, Neurosurgery & Psychiatry 52 (6), 786–787.

Faria, B. M., Duman, F., Zheng, C. X., Waisbourd, M., Gupta, L., Ali, M., Zangalli, C., Lu, L., Wizov, S. S., Spaeth, E., et al., 2015. Evaluating contrast sensitivity in age-related macular degeneration using a novel computer-based test, the spaeth/richman contrast sensitivity test. Retina 35 (7), 1465–1473.

Farsiu, S., Chiu, S. J., O'Connell, R. V., Folgar, F. A., Yuan, E., Izatt, J. A., Toth, C. A., 2014. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology 121 (1), 162–172.

Fawcett, T., 2004. Roc graphs: Notes and practical considerations for researchers. Machine learning 31 (1), 1–38.

Ferris, F. L., Sperduto, R. D., 1982. Standardized illumination for visual acuity testing in clinical research. American journal of ophthalmology 94 (1), 97–98.

Ferris, F. L., Wilkinson, C., Bird, A., Chakravarthy, U., Chew, E., Csaky, K., Sadda, S. R., 2013. Clinical classification of age-related macular degeneration. Ophthalmology 120 (4), 844–851.

Ferris III, F. L., Kassoff, A., Bresnick, G. H., Bailey, I., 1982. New visual acuity charts for clinical research. American journal of ophthalmology 94 (1), 91–96.

Fine, S. L., Berger, J. W., Maguire, M. G., Ho, A. C., 2000. Age-related macular degeneration. New England Journal of Medicine 342 (7), 483–492.

Ford, J. A., Lois, N., Royle, P., Clar, C., Shyangdan, D., Waugh, N., 2013. Current treatments in diabetic macular oedema: systematic review and meta-analysis. BMJ open 3 (3), e002269.

Forgy, E. W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics 21, 768–769.

Frenkel, R. E., Shapiro, H., Stoilov, I., 2016. Predicting vision gains with anti-vegf therapy in neovascular age-related macular degeneration patients by using low-luminance vision. British Journal of Ophthalmology 100 (8), 1052–1057.

Fujimoto, J., Swanson, E., 2016. The development, commercialization, and impact of optical coherence tomography. Investigative Ophthalmology and Visual Science 57 (9).

Gabriele, M. L., Wollstein, G., Ishikawa, H., Xu, J., Kim, J., Kagemann, L., Folio, L. S., Schuman, J. S., 2010. Three dimensional optical coherence tomography imaging: advantages and advances. Progress in retinal and eye research 29 (6), 556–579.

Gal, Y., Ghahramani, Z., 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256.

Gobburu, J., 2009. Biomarkers in clinical drug development. Clinical Pharmacology & Therapeutics 86 (1), 26–27.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680.

Götzinger, E., Pircher, M., Hitzenberger, C. K., 2005. High speed spectral domain polarization sensitive optical coherence tomography of the human retina. Optics express 13 (25), 10217–10229.

Grewal, M., Srivastava, M. M., Kumar, P., Varadarajan, S., 2018. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. IEEE, pp. 281–284.

Group, A.-R. E. D. S. R., et al., 1999. The age-related eye disease study (areds): design implications areds report no. 1. Controlled clinical trials 20 (6), 573.

Group, A.-R. E. D. S. R., et al., 2000. Risk factors associated with age-related macular degeneration: a case-control study in the age-related eye disease study: age-related eye disease study report number 3. Ophthalmology 107 (12), 2224–2232.

Group, A.-R. E. D. S. R., et al., 2001a. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins c and e, beta carotene, and zinc for age-related macular degeneration and vision loss: Areds report no. 8. Archives of ophthalmology 119 (10), 1417.

Group, B. D. W., Atkinson Jr, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., et al., 2001b. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology & therapeutics 69 (3), 89–95.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316 (22), 2402–2410.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., Seung, H. S., 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405 (6789), 947.

Hajian-Tilaki, K., 2013. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine 4 (2), 627.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. In: The elements of statistical learning. Springer, pp. 485–585.

Hayton, P. M., Schölkopf, B., Tarassenko, L., Anuzis, P., 2001. Support vector novelty detection applied to jet engine vibration spectra. In: Advances in neural information processing systems. pp. 946–952.

He, K., Zhang, X., Ren, S., et al., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of IEEE ICCV. pp. 1026–1034.

Henkind, P., Hansen, R., Szalay, J., et al., 1979. Ocular circulation. Physiology of the human eye and visual system, 98–155.

Hildebrand, G. D., Fielder, A. R., 2011. Anatomy and physiology of the retina. In: Pediatric retina. Springer, pp. 39–65.

Hinton, G., Srivastava, N., Swersky, K., 2012a. Rmsprop: Divide the gradient by a running average of its recent magnitude. Neural networks for machine learning, Coursera lecture 6e.

Hinton, G. E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18 (7), 1527–1554.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R., 2012b. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Hirooka, K., Misaki, K., Nitta, E., Ukegawa, K., Sato, S., Tsujikawa, A., 2016. Comparison of macular integrity assessment (maia™), mp-3, and the humphrey field analyzer in the evaluation of the relationship between the structure and function of the macula. PloS one 11 (3), e0151000.

Ho, A. C., Busbee, B. G., Regillo, C. D., Wieland, M. R., Van Everen, S. A., Li, Z., Rubio, R. G., Lai, P., Group, H. S., et al., 2014. Twenty-four-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration. Ophthalmology 121 (11), 2181–2192.

Holz, F. G., Sadda, S. R., Busbee, B., Chew, E. Y., Mitchell, P., Tufail, A., Brittain, C., Ferrara, D., Gray, S., Honigberg, L., Martin, J., Tong, B., Ehrlich, J. S., Bressler, N. M., 2018. Efficacy and safety of lampalizumab for geographic atrophy due to age-related macular degeneration: Chroma and spectri phase 3 randomized clinical trials. JAMA ophthalmology 136 (6), 666–677.

Hong, W. S., Haimovich, A. D., Taylor, R. A., 2018. Predicting hospital admission at emergency department triage using machine learning. PloS one 13 (7), e0201016.

Hornik, K., Feinerer, I., Kober, M., Buchta, C., 2012. Spherical k-means clustering. Journal of Statistical Software 50 (10), 1–22.

Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A., et al., 1991. Optical coherence tomography. Science 254 (5035), 1178–1181.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Hubschman, J. P., Reddy, S., Schwartz, S. D., 2009. Age-related macular degeneration: current treatments. Clinical ophthalmology (Auckland, NZ) 3, 155.

Hughes, T., Deininger, M., Hochhaus, A., Branford, S., Radich, J., Kaeda, J., Baccarani, M., Cortes, J., Cross, N. C., Druker, B. J., et al., 2006. Monitoring cml patients responding to treatment with tyrosine kinase inhibitors: review and recommendations for harmonizing current methodology for detecting bcr-abl transcripts and kinase domain mutations and for expressing results. Blood 108 (1), 28–37.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data clustering: a review. ACM computing surveys (CSUR) 31 (3), 264–323.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Vol. 112. Springer.

Joussen, A. M., Bornfeld, N., 2009. The treatment of wet age-related macular degeneration. Deutsches Ärzteblatt International 106 (18), 312.

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., Karssemeijer, N., Lillholm, M., 2016. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging 35 (5), 1322–1331.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Kaufman, P. L., Levin, L. A., Adler, F. H., Alm, A., 2011. Adler's Physiology of the Eye. Elsevier Health Sciences.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems. pp. 5574–5584.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D. P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Klein, M. L., Ferris III, F. L., Armstrong, J., Hwang, T. S., Chew, E. Y., Bressler, S. B., Chandra, S. R., Group, A. R., et al., 2008. Retinal precursors and the development of geographic atrophy in age-related macular degeneration. Ophthalmology 115 (6), 1026–1031.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al., 2005. Complement factor h polymorphism in age-related macular degeneration. Science 308 (5720), 385–389.

Koch, M., 2018. Artificial intelligence is becoming natural. Cell 173 (3), 531–533.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

Krogh, A., Hertz, J. A., 1992. A simple weight decay can improve generalization. In: Advances in neural information processing systems. pp. 950–957.

Kumaran, S. K., Dogra, D. P., Roy, P. P., 2019. Anomaly detection in road traffic using visual surveillance: A survey. arXiv preprint arXiv:1901.08292.

Lad, E. M., Mukherjee, D., Stinnett, S. S., Cousins, S. W., Potter, G. G., Burke, J. R., Farsiu, S., Whitson, H. E., 2018. Evaluation of inner retinal layers as biomarkers in mild cognitive impairment to moderate Alzheimer's disease. PloS one 13 (2), e0192646.

Lambert, N. G., ElShelmani, H., Singh, M. K., Mansergh, F. C., Wride, M. A., Padilla, M., Keegan, D., Hogg, R. E., Ambati, B. K., 2016. Risk factors and biomarkers of age-related macular degeneration. Progress in retinal and eye research 54, 64–102.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48 (4), 441–446.

Langley, P., 1994. Selection of relevant features in machine learning. In: Proceedings of the AAAI Fall symposium on relevance. pp. 1–5.

Langs, G., Menze, B. H., Lashkari, D., Golland, P., 2011. Detecting stable distributed patterns of brain activation using gini contrast. NeuroImage 56 (2), 497–507.

Larsen, M., Waldstein, S. M., Boscia, F., Gerding, H., Monés, J., Tadayoni, R., Priglinger, S., Wenzel, A., Barnes, E., Pilz, S., et al., 2016. Individualized ranibizumab regimen driven by stabilization criteria for central retinal vein occlusion: twelve-month results of the crystal study. Ophthalmology 123 (5), 1101–1111.

Lassere, M. N., Johnson, K. R., Boers, M., Tugwell, P., Brooks, P., Simon, L., Strand, V., Conaghan, P. G., Ostergaard, M., Maksymowych, W. P., et al., 2007. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. The Journal of rheumatology 34 (3), 607–615.

Le, H., Borji, A., 2017. What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks? arXiv preprint arXiv:1705.07049.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521 (7553), 436.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1 (4), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.

LeCun, Y. A., Bottou, L., Orr, G. B., Müller, K.-R., 2012. Efficient backprop. In: Neural networks: Tricks of the trade. Springer, pp. 9–48.

Lee, C. S., Baughman, D. M., Lee, A. Y., 2017. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. Ophthalmology Retina 1 (4), 322–327.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports 7 (1), 17816.

Li, H., Achim, A., Bull, D., 2012. Unsupervised video anomaly detection using feature clustering. IET signal processing 6 (5), 521–533.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., Sánchez, C. I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE transactions on information theory 28 (2), 129–137.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R., Padilla, P., Gómez-Río, M., Initiative, A. D. N., et al., 2011. Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. Neurocomputing 74 (8), 1260–1271.

Madani, A., Arnaout, R., Mofrad, M., Arnaout, R., 2017. Fast and accurate classification of echocardiograms using deep learning. arXiv preprint arXiv:1706.08658.

Malhotra, P., Vig, L., Shroff, G., Agarwal, P., 2015. Long short term memory networks for anomaly detection in time series. In: Proceedings. Presses universitaires de Louvain, p. 89.

Markou, M., Singh, S., 2003. Novelty detection: a review—part 1: statistical approaches. Signal processing 83 (12), 2481–2497.

Markowitz, S. N., Reyes, S. V., 2013. Microperimetry and clinical practice: an evidence-based review. Canadian Journal of Ophthalmology/Journal Canadien d'Ophtalmologie 48 (5), 350–357.

Masci, J., Meier, U., Cireşan, D., Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In: International Conference on Artificial Neural Networks. Springer, pp. 52–59.

Matteoli, S., Diani, M., Theiler, J., 2014. An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (6), 2317–2336.

McHarg, S., Clark, S. J., Day, A. J., Bishop, P. N., 2015. Age-related macular degeneration and the role of the complement system. Molecular immunology 67 (1), 43–50.

Medeiros, F. A., 2015. Biomarkers and surrogate endpoints in glaucoma clinical trials. British Journal of Ophthalmology 99 (5), 599–603.

Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J., 2016. Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163.

Michels, S., Hansmann, F., Geitzenauer, W., Schmidt-Erfurth, U., 2006. Influence of treatment parameters on selectivity of verteporfin therapy. Investigative ophthalmology & visual science 47 (1), 371–376.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Molina-Martín, A., Piñero, D. P., Pérez-Cambrodí, R. J., 2016. Reliability and intersession agreement of microperimetric and fixation measurements obtained with a new microperimeter in normal eyes. Current eye research 41 (3), 400–409.

Mookiah, M. R. K., Acharya, U. R., Chua, C. K., Lim, C. M., Ng, E., Laude, A., 2013. Computer-aided diagnosis of diabetic retinopathy: A review. Computers in biology and medicine 43 (12), 2136–2155.

Nair, T., Precup, D., Arnold, D. L., Arbel, T., 2018. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 655–663.

Nalejska, E., Maczyńska, E., Lewandowska, M. A., 2014. Prognostic and predictive biomarkers: tools in personalized oncology. Molecular diagnosis & therapy 18 (3), 273–284.

Nowak, J. Z., 2006. Age-related macular degeneration (amd): pathogenesis and therapy. Pharmacological Reports 58 (3), 353.

Obermeyer, Z., Emanuel, E. J., 2016. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine 375 (13), 1216.

of Age-Related Macular Degeneration with Photodynamic Therapy (TAP) Study Group, T., et al., 1999. Photodynamic therapy of subfoveal choroidal neovascularization in age-related macular degeneration with verteporfin. one-year results of 2 randomized clinical trials-tap report 1. Arch ophthalmol 117, 1329–1345.

Opitz, D. W., 1999. Feature selection for ensembles. AAAI/IAAI 379, 384.

Opstal-van Winden, A. W., Rodenburg, W., Pennings, J. L., van Oostrom, C., Beijnen, J. H., Peeters, P. H., van Gils, C. H., de Vries, A., 2012. A bead-based multiplexed immunoassay to evaluate breast cancer biomarkers for early detection in pre-diagnostic serum. International journal of molecular sciences 13 (10), 13587–13604.

Orlando, J. I., 2017. Machine learning for ophthalmic screening and diagnostics from fundus images.

Owsley, C., Sloane, M. E., 1987. Contrast sensitivity, acuity, and the perception of'real-world'targets. British Journal of Ophthalmology 71 (10), 791–796.

Oyster, C. W., 1999. The human eye. Sunderland, MA: Sinauer.

Panwar, N., Huang, P., Lee, J., Keane, P. A., Chuan, T. S., Richhariya, A., Teoh, S., Lim, T. H., Agrawal, R., 2016. Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare. Telemedicine and e-Health 22 (3), 198–208.

Pascolini, D., Mariotti, S. P., 2012. Global estimates of visual impairment: 2010. British Journal of Ophthalmology 96 (5), 614–618.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544.

Paulovich, A. G., Whiteaker, J. R., Hoofnagle, A. N., Wang, P., 2008. The interface between biomarker discovery and clinical validation: the tar pit of the protein biomarker pipeline. PROTEOMICS–Clinical Applications 2 (10-11), 1386–1402.

Peng, T., Leckie, C., Ramamohanarao, K., 2007. Information sharing for distributed intrusion detection systems. Journal of Network and Computer Applications 30 (3), 877–899.

Perlis, R., 2011. Translating biomarkers to clinical practice. Molecular psychiatry 16 (11), 1076.

Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L., 2014. A review of novelty detection. Signal Processing 99, 215–249.

Pokrajac, D., Lazarevic, A., Latecki, L. J., 2007. Incremental local outlier detection for data streams. In: 2007 IEEE symposium on computational intelligence and data mining. IEEE, pp. 504–515.

Pomerol, J.-C., 1997. Artificial intelligence and human decision making. European Journal of Operational Research 99 (1), 3–25.

Popescu, D. P., Flueraru, C., Mao, Y., Chang, S., Disano, J., Sherif, S., Sowa, M. G., et al., 2011. Optical coherence tomography: fundamental principles, instrumental designs and biomedical applications. Biophysical reviews 3 (3), 155.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., Webster, D. R., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering 2 (3), 158.

Potsaid, B., Gorczynska, I., Srinivasan, V. J., Chen, Y., Jiang, J., Cable, A., Fujimoto, J. G., 2008. Ultrahigh speed spectral/fourier domain oct ophthalmic imaging at 70,000 to 312,500 axial scans per second. Optics express 16 (19), 15149–15169.

Poultney, C., Chopra, S., Cun, Y. L., et al., 2007. Efficient learning of sparse representations with an energy-based model. In: Advances in neural information processing systems. pp. 1137–1144.

Prentice, R. L., 1989. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in medicine 8 (4), 431–440.

Prokofyeva, E., Zrenner, E., 2012. Epidemiology of major eye diseases leading to blindness in europe: a literature review. Ophthalmic research 47 (4), 171–188.

Provis, J. M., Penfold, P. L., Cornish, E. E., Sandercoe, T. M., Madigan, M. C., 2005. Anatomy and development of the macula: specialisation and the vulnerability to macular degeneration. Clinical and Experimental Optometry 88 (5), 269–281.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., Williams, S. M., 2001. Neuroscience. anatomy of the eye. MA: Sinauer Associates.

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12 (1), 145–151.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rajamanickam, M., 2007. Modern General Psychology, (revised And Expanded)(in 2 Vols.). Concept Publishing Company.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1577–1581.

Regan, D., Neima, D., 1983. Low-contrast letter charts as a test of visual function. Ophthalmology 90 (10), 1192–1200.

Ribeiro, M., Lazzaretti, A. E., Lopes, H. S., 2018. A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters 105, 13–22.

Richman, J., Spaeth, G. L., Wirostko, B., 2013. Contrast sensitivity basics and a critique of currently available tests. Journal of Cataract & Refractive Surgery 39 (7), 1100–1106.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. of MICCAI. Springer, pp. 234–241.

Roy, S., Butman, J. A., Reich, D. S., Calabresi, P. A., Pham, D. L., 2018. Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. arXiv preprint arXiv:1803.09172.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S. A., Vandermeulen, R., Binder, A., Müller, E., Kloft, M., 2018. Deep one-class classification. In: International Conference on Machine Learning. pp. 4390–4399.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning representations by back-propagating errors. nature 323 (6088), 533.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one 10 (3), e0118432.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems. pp. 2234–2242.

Savino, P. J., Danesh-Meyer, H. V., 2012. Color Atlas and Synopsis of Clinical Ophthalmology–Wills Eye Institute–Neuro-Ophthalmology. Lippincott Williams & Wilkins.

Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., Ng, A. Y., 2011. On random weights and unsupervised feature learning. In: ICML. Vol. 2. p. 6.

Schatz, H., McDonald, H. R., 1989. Atrophic macular degeneration: rate of spread of geographic atrophy anti visual loss. Ophthalmology 96 (10), 1541–1551.

Schlegl, T., Ofner, J., Langs, G., 2014. Unsupervised pre-training across image domains improves lung tissue classification. In: International MICCAI Workshop on Medical Computer Vision. Springer, pp. 82–93.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., Schmidt-Erfurth, U., 2019. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.

Schlegl, T., Waldstein, S. M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B. S., Langs, G., Schmidt-Erfurth, U., 2018. Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 125 (4), 549–558.

Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., Langs, G., 2015. Predicting semantic descriptions from medical images with convolutional neural networks. In: Information Processing in Medical Imaging. Springer, pp. 437–448.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks 61, 85–117.

Schmidt-Erfurth, U., Bogunovic, H., Sadeghipour, A., Schlegl, T., Langs, G., Gerendas, B. S., Osborne, A., Waldstein, S. M., 2018a. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. Ophthalmology Retina 2 (1), 24–30.

Schmidt-Erfurth, U., Chong, V., Loewenstein, A., Larsen, M., Souied, E., Schlingemann, R., Eldem, B., Monés, J., Richard, G., Bandello, F., 2014. Guidelines for the management of neovascular age-related macular degeneration by the european society of retina specialists (euretina). British Journal of Ophthalmology 98 (9), 1144–1167.

Schmidt-Erfurth, U., Leitgeb, R. A., Michels, S., Povazay, B., Sacu, S., Hermann, B., Ahlers, C., Sattmann, H., Scholda, C., Fercher, A. F., et al., 2005. Three-dimensional ultrahigh-resolution optical coherence tomography of macular diseases. Investigative ophthalmology & visual science 46 (9), 3393–3402.

Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., Bogunović, H., 2018b. Artificial intelligence in retina. Progress in retinal and eye research.

Schmidt-Erfurth, U., Waldstein, S. M., 2016. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Progress in Retinal and Eye Research 50, 1–24.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C., 2001. Estimating the support of a high-dimensional distribution. Neural computation 13 (7), 1443–1471.

Schrag, D., Kuntz, K. M., Garber, J. E., Weeks, J. C., 2000. Life expectancy gains from cancer prevention strategies for women with breast cancer and brca1 or brca2 mutations. Jama 283 (5), 617–624.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., Poggio, T., 2007. A quantitative theory of immediate visual recognition. Progress in brain research 165, 33–56.

Siderov, J., Tiu, A. L., 1999. Variability of measurements of visual acuity in a large eye clinic. Acta Ophthalmologica Scandinavica 77 (6), 673–676.

Sidibé, D., Sankar, S., Lemaître, G., Rastgoo, M., Massich, J., Cheung, C. Y., Tan, G. S., Milea, D., Lamoureux, E., Wong, T. Y., et al., 2017. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. Computer Methods and Programs in Biomedicine 139, 109–117.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427–437.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. Biol. Skr. 5, 1–34.

Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15 (1), 1929–1958.

Stein, J. D., Newman-Casey, P. A., Mrinalini, T., Lee, P. P., Hutton, D. W., 2014. Cost-effectiveness of bevacizumab and ranibizumab for newly diagnosed neovascular macular degeneration. Ophthalmology 121 (4), 936–945.

Sunness, J. S., Rubin, G. S., Broman, A., Applegate, C. A., Bressler, N. M., Hawkins, B. S., 2008. Low luminance visual dysfunction as a predictor of subsequent visual acuity loss from geographic atrophy in age-related macular degeneration. Ophthalmology 115 (9), 1480–1488.

Suykens, J. A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural processing letters 9 (3), 293–300.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Tadayoni, R., Waldstein, S. M., Boscia, F., Gerding, H., Pearce, I., Priglinger, S., Wenzel, A., Barnes, E., Gekkieva, M., Pilz, S., et al., 2016. Individualized stabilization criteria–driven ranibizumab versus laser in branch retinal vein occlusion: Six-month results of brighter. Ophthalmology 123 (6), 1332–1344.

Taha, A. A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging 15 (1), 29.

Taur, J.-S., Tao, C.-W., 1996. Medical image compression using principal component analysis. In: Image Processing, 1996. Proceedings., International Conference on. Vol. 2. IEEE, pp. 903–906.

Tax, D. M., Duin, R. P., 1999. Support vector domain description. Pattern recognition letters 20 (11-13), 1191–1199.

Toennies, K. D., 2017. Guide to medical image analysis. Springer.

Trobe, J. D., Beck, R. W., Moke, P. S., Cleary, P. A., 1996. Contrast sensitivity and other vision tests in the optic neuritis treatment trial. American journal of ophthalmology 121 (5), 547–553.

Vilensky, J. A., Robertson, W., Suarez-Quian, C. A., 2015. The Clinical Anatomy of the Cranial Nerves: The Nerves of" On Old Olympus Towering Top". John Wiley & Sons.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. ACM, pp. 1096–1103.

Vogl, W.-D., Waldstein, S. M., Gerendas, B. S., Schlegl, T., Langs, G., Schmidt-Erfurth, U., 2017a. Analyzing and predicting visual acuity outcomes of anti-VEGF therapy by a longitudinal mixed effects model of imaging and clinical data. Investigative ophthalmology & visual science 58 (10), 4173–4181.

Vogl, W.-D., Waldstein, S. M., Gerendas, B. S., Schmidt-Erfurth, U., Langs, G., 2017b. Predicting macular edema recurrence from spatio-temporal signatures in optical coherence tomography images. IEEE transactions on medical imaging 36 (9), 1773–1783.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R., 2013. Regularization of neural networks using dropconnect. In: International Conference on Machine Learning. pp. 1058–1066.

Wang, E., Cho, W. C., Wong, S. C., Liu, S., 2017. Disease biomarkers for precision medicine: challenges and future opportunities. Genomics, proteomics & bioinformatics 15 (2), 57.

Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C., 2010. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. Neuroimage 50 (4), 1519–1535.

Wells III, W. M., 2016. Medical image analysis–past, present, and future.

Wenick, A. S., Bressler, N. M., 2012. Diabetic macular edema: current and emerging therapies. Middle East African journal of ophthalmology 19 (1), 4.

Wiesler, S., Ney, H., 2011. A convergence analysis of log-linear training. In: Advances in Neural Information Processing Systems. pp. 657–665.

Williams, M. A., Moutray, T. N., Jackson, A. J., 2008. Uniformity of visual acuity measures in published studies. Investigative ophthalmology & visual science 49 (10), 4321–4327.

Wong, E. N., De Soyza, J. D., Mackey, D. A., Constable, I. J., Chen, F. K., 2017. Intersession test–retest variability of microperimetry in type 2 macular telangiectasia. Translational vision science & technology 6 (6), 7–7.

Wong, W. L., Su, X., Li, X., Cheung, C. M. G., Klein, R., Cheng, C.-Y., Wong, T. Y., 2014. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. The Lancet Global Health 2 (2), e106–e116.

Wu, Z., Ayton, L. N., Guymer, R. H., Luu, C. D., 2013. Intrasession test–retest variability of microperimetry in age-related macular degeneration. Investigative ophthalmology & visual science 54 (12), 7378–7385.

Wu, Z., Ayton, L. N., Guymer, R. H., Luu, C. D., 2014. Low-luminance visual acuity and microperimetry in age-related macular degeneration. Ophthalmology 121 (8), 1612–1619.

Wu, Z., Jung, C. J., Ayton, L. N., Luu, C. D., Guymer, R. H., 2015. Test–retest repeatability of microperimetry at the border of deep scotomas. Investigative ophthalmology & visual science 56 (4), 2606–2611.

Wykoff, C. C., Clark, W. L., Nielsen, J. S., Brill, J. V., Greene, L. S., Heggen, C. L., 2018. Optimizing anti-vegf treatment outcomes for patients with neovascular age-related macular degeneration. Journal of managed care & specialty pharmacy 24 (2-a Suppl), S3–S15.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057.

Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., Chen, S.-J., Dekker, J. M., Fletcher, A., Grauslund, J., et al., 2012. Global prevalence and major risk factors of diabetic retinopathy. Diabetes care, DC_111909.

Yu, D., Hung, M.-C., 2000. Overexpression of erbb2 in cancer and erbb2-targeting strategies. Oncogene 19 (53), 6115.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., Chandrasekhar, V. R., 2018. Efficient gan-based anomaly detection. arXiv preprint arXiv:1802.06222.

Zhao, J., Mathieu, M., Goroshin, R., Lecun, Y., 2016. Stacked what-where auto-encoders. arXiv preprint arXiv:1506.02351.

Zheng, Y.-J., Zhou, X.-H., Sheng, W.-G., Xue, Y., Chen, S.-Y., 2018. Generative adversarial network based telecom fraud detection at the receiving bank. Neural Networks 102, 78–86.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929.

Zhou, C., Paffenroth, R. C., 2017. Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 665–674.

Zhou, Y., Arpit, D., Nwogu, I., Govindaraju, V., 2015. Is joint training better for deep auto-encoders. arXiv preprint, arXiv: 1405.1380.

Zhou, Z.-H., Zhang, M.-L., 2007. Multi-instance multi-label learning with application to scene classification. In: Advances in neural information processing systems. pp. 1609–1616.