Artificial intelligence in retina

Ursula Schmidt-Erfurth, Amir Sadeghipour, Bianca S. Gerendas, Sebastian M. Waldstein, Hrvoje Bogunović

PII: S1350-9462(18)30011-9

DOI: 10.1016/j.preteyeres.2018.07.004

Reference: JPRR 728

To appear in: *Progress in Retinal and Eye Research*

Received Date: 30 May 2018

Revised Date: 24 July 2018

Accepted Date: 31 July 2018

Please cite this article as: Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunović, H., Artificial intelligence in retina, *Progress in Retinal and Eye Research* (2018), doi: 10.1016/j.preteyeres.2018.07.004.

Artificial Intelligence in Retina

Ursula Schmidt-Erfurth, MD[1], Amir Sadeghipour, PhD[1] , Bianca S. Gerendas, MD[1], Sebastian M. Waldstein, MD, PhD[1], Hrvoje Bogunović, Phd[1]

[1]Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Corresponding author:

Ursula Schmidt-Erfurth, MD

Professor and Chair

Department of Ophthalmology

Medical University Vienna

Spitalgasse 23

1090 Vienna, Austria

Tel: +43 1 40400 79310

Fax: +43 1 40400 79120

Email: ursula.schmidt-erfurth@meduniwien.ac.at

Financial disclosures:

Ursula Schmidt-Erfurth: consultancy Boehringer Ingelheim, Genentech, Novartis and Roche

Amir Sadeghipour: none

Bianca S. Gerendas: consultancy Roche and Novartis, research support iDx and Novartis

Sebastian M. Waldstein: consultancy Novartis, speaker fees Bayer, research support Bayer and Genentech

Hrvoje Bogunovic: none

Financial support: Austrian Federal Ministry of Science, Research and Economy

Abstract:

Major advances in diagnostic technologies are offering unprecedented insight into the condition of the retina and beyond ocular disease. Digital images providing millions of morphological datasets can fast and non-invasively be analyzed in a comprehensive manner using artificial intelligence (AI). Methods based on machine learning (ML) and particularly deep learning (DL) are able to identify, localize and quantify pathological features in almost every macular and retinal disease. Convolutional neural networks thereby mimic the path of the human brain for object recognition through learning of pathological features from training sets, supervised ML, or even extrapolation from patterns recognized independently, unsupervised ML. The methods of AI-based retinal analyses are diverse and differ widely in their applicability, interpretability and reliability in different datasets and diseases. Fully automated AI-based systems have recently been approved for screening of diabetic retinopathy (DR). The overall potential of ML/DL includes screening, diagnostic grading as well as guidance of therapy with automated detection of disease activity, recurrences, quantification of therapeutic effects and identification of relevant targets for novel therapeutic approaches. Prediction and prognostic conclusions further expand the potential benefit of AI in retina which will enable personalized health care as well as large scale management and will empower the ophthalmologist to provide high quality diagnosis/therapy and successfully deal with the complexity of 21st century ophthalmology.

Keywords:

Artificial intelligence (AI)
Machine learning (ML)
Deep learning (DL)
Automated screening
Prognosis and prediction
Personalized healthcare (PHC)

Table of Contents

1. Introduction ... 3

1.1. What is AI? ... 5

1.2. *Convolutional neural networks* ... 6

1.3. Success Stories in the medical field ... 7

2. AI technology in retina ... 10

2.1. Procedures in machine-learning ... 10

2.2. Evaluation of performance ... 11

2.3. Classification methods ... 12

2.3.1. Classic Machine Learning ... 12

2.3.2. Deep Learning ... 13

2.4. Segmentation methods ... 15

2.4.1. Classic Machine Learning ... 16

2.4.2. Deep Learning ... 17

2.4.3. *Bayesian approaches* ... 19

2.5. Prediction of clinical outcomes ... 19

2.6. Alternative Scenarios ... 21

2.7. Interpretability ... 22

3. Clinical Applications of AI in retinal disease ... 24

3.1. Automated detection and quantification of features ... 24

3.1.1. Single feature detection versus entire image classification ... 24

3.1.2. Detection and Quantification of features ... 26

3.2. Screening for Retinal Disease ... 28

3.2.1. Diabetic Retinopathy Screening ... 29

3.2.2. Screening for Age-Related Macular Degeneration ... 30

3.2.3. Glaucoma Screening ... 31

3.2.4. Retinopathy of Prematurity Screening ... 32

3.2.5. Screening for retinal disease in general ... 33

3.3. Diagnostic grading/staging of retinal disease ... 34

3.3.1. Diagnostic grading/staging of DR ... 34

3.3.2. Diagnostic grading/staging of AMD ... 35

3.3.3. Diagnostic grading/staging of Glaucoma and Retinopathy of Prematurity ... 36

3.3.4. Diagnostic grading for systemic disease ... 36

3.4. Guidance of therapy ... 38

3.4.1. Automated detection of disease activity ... 38

3.4.2. Automated quantification of pathology ... 39

3.4.3. Prediction of need for retreatment ... 40

3.5. Prediction and prognosis 43

3.5.1. Prediction of visual acuity outcomes 43

3.5.2. Prediction of future natural disease course 45

4. Discussion 48

4.1. The potential of AI 48

4.2. AI and personalized medicine 49

4.3. Challenges in AI-based retina 50

1. Introduction

No field in ophthalmology has been scientifically and clinically blessed as much as retina in recent years. Retinal disease is given intensive and widespread attention with a common understanding that the condition of the retina is among the leading causes of severe vision loss and blindness on the global level. Age-related macular degeneration (AMD) currently affects 170 million people world-wide (Pennington and DeAngelis, 2016), while diabetic retinopathy (DR) is recognized as a world-wide epidemic. A third of an estimated 285 million people with diabetes have signs of DR and one third of them have vision-threatening DR (Lee et al., 2015). Furthermore, the numbers are increasing: it is anticipated that 288 million people will have AMD by 2040 and the number with DR will triple by 2050. On the other hand, therapeutic improvements in retina count among the major break-throughs in modern medicine. The introduction of intravitreal vascular endothelial growth factor (VEGF) inhibition in 2006 hugely reduced legal blindness rates and achieved considerable improvements in vision in neovascular AMD and diabetic macular edema (DME) (Varma et al., 2015). However, the success story of anti-VEGF therapy in clinical studies comes with the bitter pill of largely inferior outcomes in the real-world setting (Mehta et al., 2018). This is mainly because of delays in identifying disease onset and progressive course, as well as the unpredictability of recurrence, which together derail long-term management, particularly in neovascular AMD, the most aggressive entity. Moreover, numerous phase II/III clinical trials in the most prevalent atrophic AMD type, which leads to irreversible loss of the central retina, have been disappointing. Even the inhibition of complement factors, believed to act as major drivers of geographic atrophy (GA), have failed to halt disease progression and vision loss, leaving the question of valid therapeutic targets and relevant biomarkers unanswered (Boyer et al., 2017). Hence, retinologists are struggling with long-term visual decline in large patient populations, health care providers face disproportionate budget drains and researchers are disheartened by failed trials.

Optimism though springs from the evolution of innovative diagnostic modalities which have developed rapidly together with therapeutic advances. Optical coherence tomography (OCT), with its non-invasive visualization of retinal structures in unprecedented resolution, is the most powerful in vivo diagnostic tool in modern medicine. Spectral domain (SD)-OCT is widely available and represents the gold standard in diagnostic imaging in the management of the leading macular diseases such as choroidal neovascularization (CNV) and DME (Schmidt-Erfurth and Waldstein, 2016). A conventional 3D OCT image is based on 20,000-52,000 A-scans per second offering a resolution of 5-7 µm (Figure 1) (Adhi and Duker, 2013). The novel swept source OCT technology has already arrived in clinical practice and provides even faster scanning with up to 100,000-236,000 A-scans per second. It also operates on longer wavelengths and allows a much faster and deeper visualization, including assessment of the choroid. OCT offers the retinologist around 60 million voxels per volume, thus providing extensive information about retinal morphology. Considering the routine work-load of a busy ophthalmological practice, it would be almost surrealistic to ask an ophthalmologist to scroll through a series of 250 B-scans for each of the dozens of retina patients examined daily, realign segmentation lines and integrate multimodal imaging sources. OCT angiography, with its high-speed and efficient algorithms allowing detection of blood flow, has made non-invasive high-resolution imaging of retinal and choroidal vasculature available to ophthalmologists in clinics and practices around the world (Spaide et al., 2017). Yet as imaging technology becomes more sophisticated, the discrepancy between imaged details and clinical interpretation grows. Even a simple marker such as central retinal thickness (CRT) does not correlate with best-corrected visual acuity (BCVA) (Browning et al., 2007). The amount of potentially relevant biomarkers is overwhelming, suggesting a multitude of different disease origins and types (Gerendas et al., 2018; Spaide, 2018). Currently, subclinical features can be visualized and identified such as hyperreflective foci (HRF), a marker not visualized by clinical ophthalmoscopy, but gaining increasing value in the prognosis of intermediate AMD and the severity of DR (Curcio et al., 2017; Fragiotta et al., 2018). The era of subclinical diagnoses has begun and a novel approach to interpretation is required.

Understanding and managing retinal disease has become vastly more complex due to the enormous accumulation of images and findings as well as all the hypotheses being put forward. Every patient appears as a “big data” challenge (Obermeyer and Lee, 2017). Obviously, the new era of diagnostic and therapeutic, scientific and clinical data manufacturing urgently requires intelligent tools to manage them adequately, safely and efficiently.

Artificial intelligence (AI) has already demonstrated proof-of-concept in medical fields such as radiology, pathology and dermatology, which have striking similarities to ophthalmology as they are deeply rooted in diagnostic imaging, the most prominent application of AI in healthcare (Figure 2) (Jiang et al., 2017). The advantages of AI in medicine are overwhelming. AI is particularly suitable for managing the complexity of 21st- century ophthalmology: it can assist clinical practice by using efficient algorithms to detect and “learn” features from large volumes of imaging data, help to reduce diagnostic and therapeutic errors and foster personalized medicine. In addition, AI can recognize disease-specific patterns and correlate novel features to gain innovative scientific insight. If ophthalmologists wish to retain control of their professional future, they will have to embrace intelligent algorithms and educate themselves to become knowledgeable in evaluating and applying AI in a constructive manner.

1.1. What is AI?

Artificial intelligence (AI) is a branch of computer science that aims to create intelligent machines. The term artificial intelligence was coined by John McCarthy, who first organized a workshop in 1956 with the goal "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". This so-called Dartmouth workshop is now considered as the birth place of AI. The branch of AI referred to as *Machine Learning* – created by Arthur Samuel in 1959 – focuses on the learning feature of intelligence by developing algorithms that extract generalized principles from data. These principles are represented as mathematical models that comprise descriptive rules of the given data. In this way, machine learning approaches formed a contrast to the other automated approaches that required the descriptive rules of the data to be defined by human experts in the field and then implemented in an automated system by computer programmers.

AI is not new to medicine. The first successful automated systems for healthcare were described as early as the 1970s. An early example was a system called MYCIN developed at Stanford University (Buchanan and Shortliffe, 1984). It was an expert system-based AI that was able to recommend appropriate antibiotics using a knowledge base composed of a large number of rules in the form of *if-then* statements. It never reached clinical practice however, but not because of weakness in its performance as it was reported to actually perform better than infectious disease specialists (Yu et al., 1979).

These first healthcare attempts were based on an AI branch called *expert systems* with the idea that *knowledge engineers* would encode the decision-making ability of clinicians into a knowledge base represented by a set of facts and rules that computers could execute algorithmically. Even though expert systems were among the first successful forms of AI (Russell and Norvig, 1995) and could operate well in limited clinical situations, the medical field and variability of pathologies proved to be so broad and complex that encoding a set of rules that would contain all the relevant clinical information was too difficult to do by hand. As a consequence, the expert systems approach was largely superseded in the 1990s by a *machine learning* branch of AI, where the "rules" would be learned by algorithms directly from a set of examples instead of being encoded by hand. Today, when we consider AI we almost exclusively have machine learning in mind.

The classic machine learning approach requires that a set of biomarkers or *features* be directly measured from the data (e.g., retinal thickness measured from an OCT image). Then, based on a training set of examples of features with known *labels*, e.g., category memberships, a *classifier* learns to recognize the correct label from the newly seen features. Once a few powerful classifiers have been developed, the effectiveness of such classic machine learning models mostly relies on the discriminative power of the chosen features which underpin the classifier performance. Thus, in classic machine learning the task of a knowledge engineer is replaced with a task of hand-engineering effective domain-specific features.

1.2. *Deep Learning and Convolutional Neural Networks*

A recurring theme in machine learning research is imitation of the neural structure of the central nervous system by creating artificial neural networks (ANNs), given that the brain is the only existing working example of a highly capable pattern recognition system. An ANN is a computing system based on a network of units called artificial neurons organized into layers. Layers of neurons perform transformations of the signal as it travels from the input (first) layer to the output (last) layer. Early ANNs from the 1990s quickly found their use in medical applications as they were recognized as good classifiers where, e.g., the input would be a set of relevant patient features and the output would be a diagnosis. They were shown to be capable of performing at the same level as an expert clinician in detecting myocardial infarction in patients presenting with anterior chest pain at an emergency department (Baxt, 1991), for diagnosing renal cancer from ultrasound (Maclin et al., 1991), or for screening of diabetic retinopathy based on features extracted from a fundus image (Gardner et al., 1996).

Even though these early forms of an ANN were outperformed by other statistical learning methods for a period of time, they were resurrected in 2012 when the new breed of *deep neural networks (DNN)* were developed. A DNN is an ANN with multiple intermediate layers positioned between the input and output layers, allowing each level to learn to transform its input signal into a gradually more abstract and higher-level representation, utilizing fewer artificial neurons than a comparable shallow ANN, making them more efficient at learning. A key benefit of DNNs is that their performance was shown to continuously improve with the size of the training dataset (Figure 3). In addition, substantial advances in computational processing power suddenly allowed such DNNs to be trained and applied within a reasonable timeframe. Thus, given enough data examples and computational power, DNN easily exceeded classic machine learning methods on standard AI benchmarks. This evolution started a new subfield of AI and machine learning called *deep learning* (LeCun et al., 2015) dedicated to exploring the capabilities of DNNs. The central idea is that a neural network, instead of just acting as a classifier, can serve as the feature extractor as well. Therefore, a single deep neural network performs both tasks and can learn to jointly extract features that are suitable for a given classification problem and to classify them. Such deep networks allow training entirely *end-to-end* because instead of learning to recognize an output category from hand-engineered features they learn to recognize it from the input signal directly (Figure 4). Thus, in deep learning, the task of hand-engineering domain-specific features is replaced by one of designing reliable deep neural network architectures.

The deep learning architecture found to be most suitable for imaging data is that of convolutional neural networks (CNNs) (LeCun et al., 1998). CNNs encode connectivity pattern between neurons that resembles the organization of the mammalian visual cortex. Such networks contain special type of layers which apply a mathematical filtering operation known as convolution, which makes individual neuron process data only for its receptive subfield and emulates its response to visual stimuli. These filters act as special feature detectors and as the input image is processed with successive convolutional layers of the network, the filters in the process get stacked together creating progressively more descriptive and sophisticated feature detectors. During training, these individual detectors are then being adjusted to detect those specific image features that are needed to solve a particular image recognition task. Trained with large annotated datasets, CNNs essentially allowed computers to start recognizing visual patterns and are primarily responsible for the recent resurgence and overwhelming interest in AI.

A significant boost in the ability of computers to recognize image content came through the ImageNet Large Scale Visual Recognition Competition (Russakovsky et al., 2015), which has been run annually since 2010. The goal was to automatically classify more than 1.2 million natural images, photographs, into 1,000 categories. By 2015, the deep learning CNN models developed were reported to reach the human level of ability at such a specific “task” of image identification (He et al., 2015). This break-through marked a new era in the perception of the role of computers for the modern world as it became obvious that AI would be able to outperform human intelligence and computers’ capacities vastly exceed those of humans in multiple scientific, medical and everyday-life tasks. In 2015, the journal *Scientific American* therefore ranked AI among the “ten big advances that will improve life, transform computing and maybe even save the planet”. With deep learning networks operating like and better than the human brain the walls fell and the paradigm-shift in AI became irreversible (Scientific-American, 2015).

1.3. Success Stories in the medical field

Deep learning-based AI has shown its abilities across multiple medical domains. It shines particularly in well-defined clinical tasks where most of the information necessary for the task is contained in the data, represented as a 1D signal (e.g., electrocardiogram), 2D or 3D medical imaging (e.g., color fundus photograph or optical coherence tomography) (Litjens et al., 2017) or structured electronic medical record. In the following, we review some of the recent prominent AI applications in medicine, overall where it has been shown to perform at least on par with medical specialists.

Applications in dermatology, due to diagnoses being based mainly on visual appearance, are especially well suited for AI. This was demonstrated by a study (Esteva et al., 2017) where the authors trained a CNN to classify lesions from photographs of skin disease. The performance of detecting malignant melanomas and carcinomas matched that of as many as 21 board-certified dermatologists combined. In another dermatological application, onychomycosis diagnosis, a CNN performed equally to or better than a panel of dermatologists with various skill levels doing the same assessment in a painstaking manual fashion (Han et al., 2018).

Radiology is another medical field where AI is expected to complement or substitute some of the visual recognition tasks performed by physicians. Recently, a CNN trained to detect pneumonia-like features or abnormalities on chest and musculoskeletal radiographs performed on par with practicing radiologists, as reported in a preprint (Rajpurkar et al., 2017b). In another application, a CNN trained to assess bone age based on pediatric hand radiographs was able to estimate age with a similar accuracy to that of a radiologist (Larson et al., 2018).

In oncology, mammography screening is expected to be supported by AI in the future. The diagnostic accuracy of a CNN system was shown in 2017 to be comparable to that of experienced radiologists (Becker et al., 2017). CNNs can also be successfully trained to be as accurate as pathologists at detecting lymph node metastases in tissue sections of women with breast cancer (Ehteshami Bejnordi et al., 2017). Treatment recommendations for breast cancer made by a commercial AI system “Watson for Oncology” have been compared in a retrospective observational study with those made by oncologists at a cancer center in India and the system showed an overall concordance of 93%, which varied by the stage of cancer (80%-97%) and patient’s age (Somashekhar et al., 2018).

In ophthalmology, automated screening based on retinal images has been a target for AI for some time (Abramoff et al., 2013; Abramoff et al., 2010). Recently, a CNN was trained to screen DR and its performance was comparable to that of a panel of ophthalmologists certified for the task (Gulshan et al., 2016). Similarly, a very high sensitivity and specificity was achieved in evaluating retinal images from large multiethnic cohorts of patients with diabetes (Ting et al., 2017). DR screening using three different commercially available software packages was not only found to be accurate, but also cost effective (Tufail et al., 2017). A CNN applied to OCT was able to successfully differentiate cases with advanced AMD or diabetic macular edema, which require timely treatment, from less severe cases (Kermany et al., 2018). The performance was equivalent to that of six ophthalmologists who made similar referrals based on the same scans. Moreover, a CNN has been trained on retinal fundus images to predict cardiovascular health risk factors such as high blood pressure and performed as well as the methods that require invasive blood tests to measure cholesterol levels (Poplin et al., 2018).

Outside of the AI applications in the imaging domain, equal success can be expected from classifying other types of signals as well. In cardiology, a CNN was trained to map a sequence of electrocardiogram samples to a sequence of rhythm classes and its performance in detecting a wide range of heart arrhythmias was (reported in a preprint) to be superior to that of board-certified cardiologists (Rajpurkar et al., 2017a). Deep learning models were able to accurately predict pathological events using a representation of patients' entire original electronic health records. Such records were successfully used to predict the need for palliative care, in-hospital mortality and unplanned readmission (Avati et al., 2017; Rajkomar et al., 2018).

Evidence of success across such a range of medical domains and applications shows that given enough training data and computing power we can design deep learning systems that match or exceed human capabilities at narrowly specified medical tasks. Thus, there is a clear opportunity to develop AI-based automated systems to read medical images (Obermeyer and Emanuel, 2016). However, the success of these deep learning methods strongly depends on the availability of large and curated datasets, which only became available recently due to the proliferation of digital imaging and data bases in medicine. For example, in the above mentioned dermatology application, 18 public data sets were used and more than 100,000 training images, two orders of magnitude larger than what was reported in the literature as used before (Esteva et al., 2017). Radiographic studies relied on NIH Chest X-Ray-14 datasets interpreting more than 100,000 frontal images from over 30,000 patients (Wang et al., 2017b). In ophthalmology, AI for multiethnic screening was trained on around 100,000 images, which is a stark contrast to the early DR screening work in the 1990s using classic machine learning training on only 300 images (Gardner et al., 1996; Ting et al., 2017).

2. AI technology in retina

To understand and properly interpret AI-based diagnostic results, the retina expert needs to become aware of the large spectrum of machine and deep learning methods which will be the bases of clinical management decisions. Diagnostic decisions of the pre-AI era were based on commonly accepted pathological clinical features and definitions of healthy versus diseased retina such as in the Early Treatment Diabetic Retinopathy study (ETDRS) or Age-Related Eye Disease Study (AREDS) scores. AI methods are also based on feature discrimination. However, the sensitivity and accuracy of referring to features as healthy or pathological/present or absent often proceeds on a subclinical base for biomarkers not seen ophthalmologically, or from integrated patterns over large populations as opposed to an individual condition. Hence, AI represents a major paradigm shift in retinal diagnosis which is unlike any previous approach. The community has to understand the rules and risks of the different methods to properly use the machine-based outputs in their daily management decisions and be aware of their reliability. Otherwise, retinology will become dependent on a “black box,” with all its inherent risks and errors, and detached from evidence-based medicine.

2.1. Procedures in machine-learning

Ophthalmology is particularly well positioned to reap the benefits of advances in AI. The clear media of cornea and lens offer easy and non-invasive diagnostic access to important functional and morphological units such as the retina and optic nerve. In practice, diagnoses has relied heavily on the use of digital imaging, starting with 2D color fundus photography (CFP) in the 1990s and then 3D OCT in the 2000s. Both modalities allow non-invasive and fast high-resolution imaging of the retina, including the retinal vasculature and the neurosensory layers. Thus, imaging has become part of clinical routine and large imaging datasets can be assembled readily. The growing number of images is one of the main driving forces behind the different applications of AI in the field of retina imaging.

Different approaches in machine learning are aimed at solving two types of learning tasks: supervised and unsupervised learning. In supervised learning, the task is to construct a mathematical model that maps given input data to their desired output values, i.e. the pathological features are a priori known and well defined. For example, from a diseased OCT volume as the input, the desired output would be the annotation of the fluid in the retina. Given CFP as the input, the desired output could be whether the retina is clinically healthy or diseased. Machine learning algorithms are designed to learn the relation between defined image features and the expert annotations supplied at an image or pixel level as ground truth. The learning process starts from a training phase, where the model is formed iteratively by the dataset (Figure 3). Once the training is completed, in a test phase, the learned model can be used to make decisions about the output of new data samples. As applying labels to the input data as ground truth for a classification task is challenging and even infeasible in some applications, there are also variations of weakly supervised learning that can deal with partly unlabeled data or noisy labels. On the other hand, in unsupervised learning the given data are completely unknown and unlabeled and the task is to create a mathematical model that describes the structure of the input data de novo.

The application of AI, and in particular machine learning in retina images, is dominated by supervised learning tasks. We identify three principal use-case scenarios for such applications (Figure 5):

Classification. To assign an image to different categories, e.g., by disease type or disease stage. Typically used for automated diagnosis, screening or staging.

Segmentation. To detect and delineate anatomical structures or lesions in an image for the purpose of measuring, e.g., shape or volume. Typically used for automated quantification of imaging biomarkers.

Prediction. To predict future outcomes or to predict the value of another measure, e.g., visual acuity, age or blood pressure. Typically used for disease prognosis or structure-function correlation.

2.2. Evaluation of performance

In supervised learning, the performance of a learned model can be evaluated based on its prediction accuracy on separate test data samples which were not present in the training dataset. If the performance of a model is strong on the training data but poor on the test data, the model has learned very specific patterns and is referred to as “overfitted” to the training data. By contrast, an underfitted model performs poorly on both training and test data. A well-fitted model performs accurately on the training data and generalizes well to the unseen data samples, which serves as an important confirmation of the outcome validity.

Different statistical measures are available to quantify the performance of a learned model with respect to data. In a classification task with positive and negative labels (e.g., result of a screening), *sensitivity* measures how many positive samples were correctly identified and *specificity* measures the proportion of correctly identified negative samples. Higher sensitivity comes at the cost of lower specificity and vice versa in many classification problems. The graphical plot of a receiver operating characteristic (ROC) curve is used to find a trade-off between them. ROC curves make use of the fact that many classification methods generate probabilities of assigning an input data sample to each possible output label. By changing the probability threshold for a classification decision, the proportion between the positive and negative label outputs change and, in this way, either the sensitivity or specificity of the model increases. An ROC curve visualizes all possible combinations of sensitivity and specificity of the model to choose from for the desired application. However, in order to measure the overall performance of a model, independent of a specific threshold and application, the area under the ROC curve (AUC) is used. The value of AUC lies between 0.5, which corresponds to a random guess, and 1.0, which shows 100% of specificity and sensitivity. Such vocabulary as ROC curve, and particularly AUC, represent important outcome endpoints in AI-based diagnosis and prognosis making.

2.3. Classification methods

The scenario targeted most by machine learning methods is image classification, i.e., the task of visual recognition consisting of assigning a particular category to an image or a volumetric scan. It is typically used in retinal analysis for the purpose of automated screening, which is an example of a binary classification: referable/non-referable or diagnosis, which is an example of a multi-class classification: type of disease present or stage of disease.

Depending on the size of the dataset available and the level of interpretability needed, there are currently two main approaches to retinal image classification: deep learning and classic machine learning. The first is the method of choice when large annotated datasets are available and when the need for higher accuracy prevails over the need for interpretability. The second is used when the annotated datasets are of small size or when there is a strong need for model transparency and interpretation of its performance.

2.3.1. Classic Machine Learning

In the classic machine learning paradigm, the focus is on creating a step-wise pipeline where the first step consists of describing an image with a set of hand-designed features. The discriminative feature set obtained is then exploited in the next step by a classifier, typically in the form of an ANN, or other statistical learning models called a support vector machine (SVM) or random forest. The performance thus relies mostly on how discriminative the features are and less on the capabilities of the classifiers, which are in practice similarly powerful. There are a few ways such features are created from an image as shown next.

Visual descriptor-based. A powerful approach for describing image content is using a *bag-of-words* model where a dictionary containing representative visual words is first created. Based on this dictionary, an image can then be characterized as features measuring the number of times such visual words appear in an image. Using such an approach, a good automated classification of different AMD stages from OCT of about 1,000 patients has been achieved (Venhuizen et al., 2017). As opposed to learning a visual dictionary, another approach relies on using known effective descriptors of local image content. One such powerful descriptor called the *histogram of oriented* *gradients* (HOG) counts occurrences of local intensity changes in different portions of an image. Such image representation was shown to be very successful in detecting objects in natural scenes and was adapted for classification of retinal OCT images (Srinivasan et al., 2014).

Segmentation-based. The availability of segmentation methods allows extraction and delineation of structures of interest from the original image. Then, descriptors can be applied on the segmented images to obtain features that quantify imaging biomarkers or lesion characteristics. Retinal layer segmentation methods allow thickness maps of different layers to be created. The thickness maps can be further summarized with a set of values by, e.g., computing a mean thickness over spatial areas defined by cylinders of various radii centered at the fovea. Such an approach was used to create a model to detect patients with AMD from OCT scans using maps of retinal volume and abnormal retinal pigment epithelium (RPE) drusen complex thickening and thinning (Farsiu et al., 2014). Similarly, Liu et al. proposed glaucoma detection from OCT based on retinal nerve fiber layer (RNFL) thickness values of circumpapillary and macula regions (Liu et al., 2011b). In another work, after optic disc detection and vessel segmentation in CFP, a set of different features based upon the color, texture, vascular and disc margin obscuration properties were extracted to capture possible changes in the optic disc in an effort to detect papilledema (Fatima et al., 2017). All these approaches highlight the accuracy of diagnostic evaluation using machine learning.

2.3.2. Deep Learning

The deep learning-based approach for retinal image classification is dependent on employing and training a CNN classification model. The classification scenario has a unique advantage because of substantial prior research done by the deep learning community in the domain of solving a related visual recognition challenge such as prompted by the ImageNet. The ImageNet is a large visual database designed for use in visual object recognition software research such as recognizing cats, dogs, or human individuals based on consistent object labelling. As of 2016, over 14 million images had been manually annotated into more than 20,000 different categories. This resulted in the availability of a few off-the-shelf, effective image classification architectures, typically coming from the past winners of the ImageNet challenge, known by the names of AlexNet, Inception, VGGnet, ResNet, etc. Thus, reuse of such CNN models, i.e., taking a known CNN architecture and initialize its variables with the ones obtained from its training on the large ImageNet dataset, has become common practice. Such a setting when the network is first *pretrained* on a different, but related, dataset is known as *transfer learning*. Alternatively, if a task-specific architecture is developed then the network variables are initialized randomly and the training is performed *from scratch*.

Transfer Learning. The idea behind transfer learning is to exploit knowledge obtained by learning to solve a related task where the training data is plentiful to allow quicker learning of the target task. ImageNet is a typical example of where the visual recognition task is used with abundant training data and is a very good model for recognizing natural photographs. Consequently, it is considered to be a good starting point for obtaining a model that performs well in recognizing retinal images. The simplest setting consists of using an already trained CNN as a fixed feature extractor. In this setting, a well-known successful CNN trained on ImageNet is typically used. Once applied to the retinal dataset, another classifier is then learned to perform the classification from such CNN-encoded features. This approach has been used to classify a central OCT B-scan into one of the four categories, CNV, early AMD, DME or normal retina, using only a fraction of data as would have been needed normally (Kermany et al., 2018). A similar approach successfully detected neovascular AMD from a conventional central OCT B-scan (Treder et al., 2018). Other researchers trained a classifier using the extracted CNN feature vector to classify fundus images into ten categories, normal retina and nine retinal diseases, using an open database containing 400 images, only a tiny fraction of the data that would normally be needed to reliably identify morphological features de novo (Choi et al., 2017).

Instead of using a CNN as a fixed feature extractor, the weights of the pretrained network can additionally be *fine-tuned* by continuing the training process on a dataset from the target domain. An example is the highly effective solution proposed for DR screening, where researchers additionally fine-tuned their model data set using ≈ 100,000 CFP images (Gulshan et al., 2016). Karri et al. fine-tuned their CNN to classify an OCT B-scan into DME, dry AMD or no pathology (Karri et al., 2017). However, fine-tuning is only successful if sufficiently large training data are available from the target domain.

Training from scratch. A customized CNN for DR screening was proposed and trained on 75,000 publicly available CFP images, where an additional classifier was further employed on the CNN-derived features to achieve the final diagnosis (Gargeya and Leng, 2017). Diagnosing AMD from CFP was proposed in the form of a custom CNN trained on 130,000 AREDS images where the task was to distinguish the disease-free or early AMD from the referable intermediate or advanced AMD. In an effort to screen a multiethnic population, researchers used almost 0.5 million images for training and validation of the task of detecting referable DR, referable possible glaucoma or referable AMD (Ting et al., 2017). Another model for grading of DR from CFP was proposed where one CNN was used for automated staging and another to provide prognosis and suggest treatment (Takahashi et al., 2017). Detecting rhegmatogenous retinal detachment from 2D ultra-widefield fundus images using a custom CNN has also been proposed (Ohsugi et al., 2017). Detection of referable AMD from CFP images was trained and evaluated on 130,000 stereo images from 4,613 patients forming the AREDS data set (Burlina et al., 2017). Researchers developed a CNN trained on more than 100,000 of the central eleven B-scans from an OCT that was able to distinguish normal from AMD retinas (Lee et al., 2017a).

A successful hybrid combination of deep learning and classic machine learning was proposed for automated screening of DR from CFP in 2016 (Abramoff et al., 2016). First a set of CNN-based detectors were applied to find the optic disc and the fovea as well as lesions such as hemorrhages, exudates and neovascularization. Features formed from these detectors were then merged and supplied to the classification stage that outputs a clinical likelihood of referable DR.

2.4. Segmentation methods

In computer science, *image segmentation* refers to the process of dividing an image into segments or outlined groups of pixels that represent a meaningful entity. Image segmentation has a wide spectrum of applications from object detection and face recognition to delineating anatomical structures in medical images. The rise of automated segmentation methods in retinal imaging first happened in the early 2000s, starting with segmentation targets on fundus images displaying features such as retinal blood vessels, microaneurysms, optic disc:cup ratio and drusen, followed by the segmentation of hemorrhage, exudates and detection of the fovea. In the 2010s, 3D OCT devices were already wide-spread and a new body of literature on automated segmentation of retinal structure from OCT volumes appeared. OCT offered insights into the retinal layers and the neurosensory layers were the primary and first segmentation targets in OCT images. Later, with the development of deep learning, more successful segmentations were reported for intra- and subretinal fluid (IRF and SRF), and more recently for drusen, pigment epithelial detachment (PED), geographic atrophy (GA), hyperreflective foci (HRF), subretinal hyperreflective material (SHRM) and photoreceptors. Figure 6 shows that the performance of the recent automated segmentations is comparable to that of human graders.

AI approaches, and in particular machine learning methods, are not essential for automated segmentation. Many structures in fundus and OCT images can be segmented by using image processing methods that solely apply a set of mathematical functions on the content of an image. By contrast, machine learning methods precisely learn those functions from annotated images. Most of the methods proposed for automated localizing and segmenting the optic disc and fovea on fundus images have not applied machine-learning methods and thus do not count as AI applications in retina imaging. For instance, Kao et al. localized the optic disc by applying functions that highlight the pixels with a yellowish hue, while the fovea is detected from a green hue of the pixels (Kao et al., 2014). In addition, the disc-fovea axis is determined guiding the foveal position towards the biologically plausible region. Such image processing methods are suitable for segmenting targets with the aid of a well-defined shape, color, texture as well as biological constraints. Most of the methods proposed for automated layer segmentation in OCT images do not apply machine learning methods either and the methods applied rely on constraints such as the ordering and thickness of the layers. Song et al. showed that when biological shape and context priors are used in a graph-based method to segment OCT layers, the segmentation error becomes significantly smaller than the corresponding inter-reader variability (Song et al., 2013). Miri et al. proposed a method to segment Bruch's membrane of patients with glaucoma using both machine learning techniques and a graph-based optimization method (Miri et al., 2017). Graphs are mathematical tools that can be seen as a network of nodes and edges that connect these nodes. For the purpose of layer segmentation, the biological distance constraints can then be encoded in the edges between nodes and in this way arrange for the most plausible positioning of the nodes. Non-AI methods have also been successfully applied to segment PED in OCT (Shi et al., 2015), hard and soft exudates (Kaur and Mittal, 2018), or hemorrhage and microaneurysms in CFP (Figueiredo et al., 2015).

In contrast to these image processing techniques, machine learning methods are not designed to work with predefined rules. This makes them more powerful in segmenting less structured targets such as cystoid fluid within the retina. The classic machine learning approaches work with numerical features that need to be extracted from given images. Features such as the relative position to a biological reference or the relative contrast of each pixel to its neighbors provide helpful information for a machine learning method to learn decision rules, whether a pixel belongs to the target segment or not. The more modern end-to-end deep learning methods work rather with raw images and do not need to be provided with such prior computed features because these are extracted implicitly during learning.

2.4.1. Classic Machine Learning

Since the 2000s, classic machine learning methods such as SVMs, Bayesian approaches and random forest (RF) have been massively applied to retina-related tasks. Each of these methods has its own strengths and weaknesses that make each of them more suitable for specific research questions and segmentation targets. Researchers and clinicians need to recognize these strengths and weaknesses to be able to make proper evaluations.

Segmentation is defined as classification of each pixel by asking questions about features such as its position and color individually and in relation to the neighboring pixels. In classic machine learning, these features are extracted from the images in a pre-processing step before applying the learning method. Random forest approaches build decision trees based on the features extracted to classify each pixel. Applications of random forest in CFP have reached performances comparable to manual grading of drusen (van Grinsven et al., 2013), pseudodrusen (van Grinsven et al., 2015), exudates (Liu et al., 2011a) and geographic atrophy (Feeny et al., 2015). In all these applications, the classification relies mostly on the color and texture features. Lang et al. applied RF to segment eight layers in OCT volumes, mainly based on the position of each pixel in the volume and its gray value (Lang et al., 2013).

Features extracted solely at the pixel level do not take the contextual information into account. Thus, different techniques are applied to incorporate the neighboring pixels in the classification decision. For instance, Wang et al. proposed an RF method to distinguish between preserved and disturbed ellipsoid zones in the en-face view of OCTs (Wang et al., 2018). The decisions per pixel were made not only based on the intensity of each isolated pixel but also on the functions of the pixel value relative to the neighboring pixels, an image processing technique known as Kernel operation. Montuoro et al. proposed the auto-context loop for a joint segmentation of retina layers and fluid (Montuoro et al., 2017). The idea is to use the probabilistic segmentation result (which carries the contextual information) as input and repeat the classification process. Another technique of context-sensitive segmentation uses image patches, which are random samples of small regions of the image. For instance, Ren et al. extracted patches with and without drusen from CFP images (Ren et al., 2018). These were then used to learn features present in drusen segments and absent in background segments. An SVM applied to segment drusen based on both learned features from patches and hand-crafted features from pixels achieved a relatively high accuracy in public datasets. The strength of SVMs is that the learning method can mathematically set boundaries on the error rate and avoid over-fitting to the training data, which is a typical problem in machine learning applications. In CFP, different SVM approaches have reliably segmented different targets such as microaneurysms and vessels. Veiga et al. proposed using an SVM in two classification phases: first the SVM classifies based on the pixel-level features and second the SVM decides at the more contextual level, based on the groups of candidate pixels from the first SVM (Veiga et al., 2017). Relan et al. also applied an SVM to automatically segment and classify vessels on CFP images into arterioles and venules (Relan et al., 2014). The features were computed as the mean of pixel color values in a neighborhood around each pixel for a context-sensitive segmentation.

2.4.2. Deep Learning

The number of deep learning applications in retinal segmentation tasks has increased in the last years. Deep learning has not only achieved comparable or better results to the previous methods of classic machine learning for any segmentation target but has also been applied to segment new targets such as SHRM and GA in OCT (Ji et al., 2018; Lee et al., 2018b). Al-Bander et al. applied deep CNNs to detect both the fovea and optic disc on color fundus images (Al-Bander et al., 2018). In contrast to the classic machine learning applications, this CNN method works directly on the raw images and does not require any prior knowledge about the morphology and position of the fovea and optic disc. Another example of CNN applications on fundus images is the segmentation and classification of arterioles and venules (Welikala et al., 2017). A similar CNN architecture has been used to segment the optical nerve head in OCT scans (Devalla et al., 2018). Both these methods take the raw image as input and process it into the desired labels per pixel in a long (deep) sequence of mathematical operations.

Schlegl et al. proposed an encoder-decoder architecture to segment both sub- and intraretinal fluid in OCT images of patients with neurovascular AMD, DME or retinal vein occlusion (RVO) (Schlegl et al., 2018b). The encoder part of the network maps the given B-scan of an OCT volume to an abstract representation with lower resolution than the image. The subsequent decoder part of the network then generates the segmentation mask (i.e. the delineated annotation) in the original resolution of the input image. By training such a model on a given set of annotated B-scan samples, the encoder part of the model learns to keep the most important information (or implicitly extracted features) from different samples so that the decoder part can generate annotations solely from this information.

Recently, other deep learning architectures have been proposed that use a composition of neuronal networks different from the mere sequential processing. One of the most successful architectures in medical image processing and also in retina imaging is U-Net (Ronneberger et al., 2015). This architecture arranges the neuronal layers in a U form. Additional connections between encoder and decoder layers allow images to be processed at different levels of abstraction. U-Net architecture has been successfully applied for segmentation in OCT scans such as of drusen (Zadeh et al., 2017), intraretinal fluid (Venhuizen et al., 2018), macular edema (Lee et al., 2017b), retinal layers (He et al., 2018) and hyperreflective foci (Schlegl et al., 2018a). Roy et al. proposed ReLayNet, a deep learning network architecture inspired by U-Net with a modified connection between the layers for segmentation (Roy et al., 2017). ReLayNet achieved accurate results in segmentation of seven retina layers together with fluid in pathological OCT scans. More recent deep learning architectures have segmented some less explored biomarkers of OCTs. Lee et al. proposed a U-Net architecture that simultaneously segments several relevant lesions of neovascular AMD (IRF, SRF, PED, and SHRM) on

OCT scans, with a sensitivity of at least 0.97 (Lee et al., 2018b). In summary, a large spectrum of deep learning methods has to be distinguished and each method performs differently.

2.4.3. *Bayesian approaches*

Bayesian methods are a family of probabilistic methods with the advantage that the mathematical model learned can be represented as a graph. This makes the model structure interpretable for humans, in contrast to SVMs or neuronal networks which represent the learned classification model in a black-box manner. Kharghanian et al. reported comparable vessel segmentation accuracy in CFP images when applying an SVM versus a Bayesian probabilistic method (Kharghanian and Ahmadyfard, 2012). Zheng et al. proposed a hybrid approach which combines Bayesian with graph-based methods to segment retinal layers in OCT (Zheng et al., 2013). This work also applied a meta-learning technique called adaptive boosting or AdaBoost that combines a set of weak classifiers with a unified boosted and strong classifier. This technique has been successfully applied to segment PED in OCT images or to segment vessels in color fundus images (Memari et al., 2017; Sun et al., 2016). Another similar technique is called ensemble learning, in which the final classification decision is made based on the votes of an ensemble of simple classifiers. For instance, Harangi et al. showed that in segmentation of exudates in CFP, an approach that ensembles a set of simple and weak Bayesian classifiers outperforms the state-of-the-art methods that apply a single but strong classifier (Harangi and Hajdu, 2014).

Besides the aforementioned supervised classification methods, successful application of unsupervised clustering methods has been reported for different segmentation targets. Methods such as k-nearest-neighbor (KNN) do not need any annotated data as ground truth and thus the extracted features from the images need to be descriptive enough for the pixels belonging to the segmentation target to have a similar range of values. KNNs have been successfully applied to segment vessels (GeethaRamani and Balasubramanian, 2018), exudates (Allam et al., 2017) and microaneurysms in CFP (Walter et al., 2007).

2.5. Prediction of clinical outcomes

Similar to a classification scenario, AI can be applied to predict completely different attributes of the patient or the future outcome of a treatment from an image. Poplin et al., using the retina as a window to the body, successfully trained a CNN on data from 284,335 patients to “guess” age, sex and systolic blood pressure from a CFP (Poplin et al., 2018). Of note, this work was published as a non-peer reviewed preprint. The task was achieved by training for multiple predictions simultaneously, referred to as multi-task learning. Having to predict multiple attributes simultaneously was shown to benefit the learning process as by sharing representations between related tasks the model is able to generalize better on the original task of interest. Prahs et al. used a total of 183,402 retinal OCT B-scans to train a CNN to predict from a central B-scan whether an anti-VEGF injection would be given in the following 21 days to a patient with neovascular AMD (Prahs et al., 2018). Longitudinal datasets are often needed in addition to make predictions because not only the current state of the retina but also its recent morphological development and change over time has to be observed (Figures 7 and 8).

Classic Machine Learning. Prediction of disease recurrence from longitudinal OCT in patients with RVO after anti-VEGF initiation has been proposed (Vogl et al., 2017b). The classifier was trained on 247 patients from spatio-temporal features measuring local retinal thickness values and their change during the first three monthly visits. Prediction of visual acuity after a period of anti-VEGF treatment from a set of spatio-temporal OCT features and clinical biomarkers was proposed for neovascular AMD (Schmidt-Erfurth et al., 2018a), DME (Gerendas et al., 2017) and RVO (Vogl et al., 2017a). There, the OCT biomarkers corresponding to the retinal layer thicknesses, volume and area covered by the retinal fluid were spatially described by their mean ETDRS grid values. Similarly, Bogunovic et al proposed prediction of anti-VEGF treatment requirements in the following two years from a set of spatio-temporal OCT biomarkers obtained during the initiation phase and showed that the automated prediction performance was comparable to or even better than that of a clinician (Bogunovic et al., 2017b).

Predictions can also be made for a local region of the retina only. Niu et al. presented prediction of GA growth on OCT (Niu et al., 2016). First, the GA area was identified. Then, the surrounding en-face pixels and their axial scan properties (e.g. mean reflectivity) and segmentation-based properties (e.g. drusen height and the presence of pathologies such as reticular pseudodrusen or loss of photoreceptors), were used to train a classifier to predict for each pixel whether it would be affected by GA at the next follow-up visit (Figure 9). A similar approach was used for predicting the regression of individual drusen from OCT biomarkers in eyes with early/intermediate AMD (Bogunovic et al., 2017a) (Figure10). There, confluent drusen were first partitioned into individual ones and prediction for each druse was performed independently from a set of spatio-temporal features describing drusen morphology, reflectivity and their surrounding layers. In glaucoma, vision is traditionally measured by Humphrey 24-2 visual field sensitivity thresholds, which is subjective and time-consuming. Bogunovic et al. and Guo et al. instead regressed local visual sensitivity from a sequence of retinal RNFL and GCL thickness sector values following a spatial connectivity model of optical nerve fibers from a wide-field OCT protocol (Bogunovic et al., 2014; Guo et al., 2017).

Survival analysis. The task of predicting the time to a future “event” relies on survival models, where the retina is considered to have “survived” until the event occurs (Figure 11). Such models estimate a risk of an event occurring and have to specifically account for the *censoring* phenomenon, i.e., the fact that only some retinas experience the event for the duration of the study or are lost to follow-up. The Cox proportional hazards (CPH) model is the most commonly used model for survival data and effectively accounts for different individual intervals and censoring. In a study by Chiu et al., the event was defined as the occurrence of the first incidence of advanced AMD in an eye (Chiu et al., 2014). The authors built an eye-specific predictive model for developing advanced AMD from CFP images of 4507 participants with AREDS. The authors used the baseline predictors, age, sex, education level, race and smoking status, and the presence of pigmentary changes, soft drusen, and maximum drusen size to devise and validate a macular risk scoring system. A set of OCT measurements describing outer retina and drusen and their change during follow-up was used to build a predictive model of the onset of neovascular AMD (de Sisternes et al., 2014). A longitudinal dataset of five consecutive follow-up visits was used to predict the risk of conversion to neovascular AMD or GA from a set of quantitative spatio-temporal OCT imaging biomarkers (Schmidt-Erfurth et al., 2018b), as shown in Figure 12.

2.6. Alternative Scenarios

In addition to the three main scenarios already covered, deep learning combined with big training datasets allowed development of applications that had not been considered previously. Two typical such applications are enhancement or restoration of already acquired retinal images and synthesizing retinal images of a different modality.

Image enhancement. An example is a deep learning method for removing blurring artifacts from adaptive optics (AO) images (Fei et al., 2017). The method learned to map between the blurred and restored retinal images in an end-to-end fashion. The mapping was represented as a deep convolutional neural network that was trained to output high-quality images from blurry inputs. The CNN was trained on 500,000 retinal image pairs with simulated optical aberrations of the eye.

Image synthesis. Lee et al. showed how to train a CNN to generate OCT angiography-like en-face images from a structural OCT image alone (Lee et al., 2018a). A total of 401,098 cross-sectional pairs of structural and corresponding angiography OCT images from 873 volumes were used as a training set. The model was able to learn and enhance the weak signal patterns in structural OCT that correspond to the vascular structures using the corresponding OCT angiography image as a guide. This trained model was able to significantly outperform clinicians in detecting vascular structures on structural OCT. However, it must be acknowledged that AI cannot uncover information that is not present in an image, as for instance flow information from structural OCT. Nevertheless, AI is able to use cues that human observers cannot routinely consider or find difficult to identify: for example, hallmarks of the retinal vasculature on OCT sections, which are well visible but often not considered systematically by physicians.

Anomaly detection. Deep learning can also be applied in an unsupervised way to discover anomalies in image data. For this purpose, the appearance and variability of normal images of healthy individuals are learned. If the AI system is then presented with an image containing disease features, these features can be recognized automatically because they appear as different from the learned healthy ones. Thus, pathological features can be identified without defining them a priori. Anomaly detection using unsupervised learning widely opens the horizon for an unbiased discovery of hitherto unknown biomarkers. The markers discovered can then be analyzed further in unsupervised cluster analyses to reveal typical pathophysiological patterns common to anomalies, i.e. defined pathological features. Successful applications of unsupervised AI systems using OCT images were recently presented by Seeböck and Schlegl (Schlegl et al., 2017; Seeboeck et al., 2016).

2.7. Interpretability

Having AI perform at an expert level is often insufficient if it operates as a black-box model, i.e., without information on how the AI model reached its decision. In classification scenarios, real-word trust in AI's performance and detection of possible model biases is built when physicians understand which discriminative features were used in the decision-making process. In prediction scenarios, we are interested in learning from AI by understanding the role of individual predictive factors and to advance our clinical insight of the underlying pathophysiology beyond conventional knowledge.

Classic machine learning. In classic machine learning, as we are building a pipeline of individual components, each stage is hand-designed and hence more interpretable. When a linear predictive model is used, weights associated with each feature often serve as a surrogate measure of its importance. This has been done to identify risk factors for conversion to advanced AMD or understand what separates anti-VEGF responders from non-responders (Bogunovic et al., 2017a; de Sisternes et al., 2014; Schmidt-Erfurth et al., 2018b; Vogl et al., 2017b). In a non-linear random forest classifier, the individual feature importance relies on permuting the values of a feature and measuring how much such permutation decreased the prediction accuracy of the model. Important features can then be detected as those where the permutation decreases the prediction accuracy most. This has been performed to understand the prediction of GA progression (Niu et al., 2016), find predictive features of treatment requirements in anti-VEGF therapy and identify predictive factors for BCVA outcomes of intravitreal anti-VEGF or conversion to late AMD Figure 12) (Bogunovic et al., 2017b; Schmidt-Erfurth et al., 2018a; Schmidt-Erfurth et al., 2018b).

Deep learning. With the growing success of neural networks, there is a corresponding need to be able to explain their decision base. However, we have to accept that deep models use greater abstractions and tighter integration at the cost of lower interpretability. The most prevalent trend in the field of neural network interpretability is to study what part of an example image is responsible for the network activating in a particular way (e.g., giving a positive or negative diagnosis). This is typically represented in the form of an image *heatmap* indicating which local morphology changes would modify the network predictions. The most common and simplest approach is to perform the so-called occlusion test (Zeiler and Fergus, 2014) applied for AMD detection on OCT (Kermany et al., 2018; Lee et al., 2017a). To identify the areas in the image contributing most to the neural network's decision, a blank box is moved across all positions in the image and the respective output class probability recorded. The highest drop in the class probability will represent the region of interest with the highest importance. Recently, a convolutional visualization layer was implanted at the end of the network to highlight prognostic regions of the fundus for DR diagnosis (Figure 13) (Gargeya and Leng, 2017). Class saliency maps (Simonyan and Zisserman, 2014) were used (Poplin et al., 2018) to highlight parts of the fundus image which were the most discriminative for the CNN when predicting individual sex, age and blood pressure (Figure 14). This approach greatly facilitates clinical understanding.

3. Clinical Applications of AI in retinal disease

In the following section, we identify clinical scenarios that are accessible by AI applications, and summarize the current state-of-the-art in AI research in each scenario. This includes automated detection and quantification of retinal lesions or features, automated screening for retinal disease, AI-based diagnostic grading as well as clinical decision support in retinal therapy and prognostic disease models.

3.1. Automated detection and quantification of features

The most common application of AI methods in retina is its use for detection of disease-related features on CFP images. An important first step for evaluating a CFP image for automated analysis is to identify if the orientation of the image is adequate for the automated system to analyze the retinal condition. The retinal landmarks mainly used for this task are the large retinal vessels and optic disc and sometimes also the foveal location because these can be found equally in every fundus image. Figure 15 shows examples of such orientational key structures (Moccia et al., 2018; Molina-Casado et al., 2017). These landmarks allow an algorithm to create a common space where every image received can be brought into the correct context – as the human observer does when performing a slit-lamp or ophthalmoscopic examination.

In OCT, the detection of the fovea is of great importance as it can also serve for orientation, particularly in macular disease (Liefers et al., 2017). The spatial context allows clinical conclusions: a structure in the perifoveal area is most likely less clinically relevant, for example, than a pathological alteration directly in the center of the fovea.

3.1.1. Single feature detection versus entire image classification

The ETDRS classification system has been broadly used for decades both in clinical classification and in the context of randomized clinical trials. It is based on the detection of retinal markers seen by slit-lamp examination or on CFP. Accordingly, most work on lesion detection in CFP has been published in DR. Microaneurysms (Wang et al., 2017a), hemorrhage (Figure 15) (van Grinsven et al., 2016) and hard exudates (García et al., 2009a; Yu et al., 2017) are the most relevant features but detection of blood vessels and their alterations, e.g., venous beading or intraretinal microvascular abnormalities (IRMA), is also important. Good image quality is vital, especially for small features, as poor quality or the wrong field of view can lead to feature omission (Wang et al., 2017a), comparable with slit-lamp examination when the ability to recognize the retinal fundus precludes the diagnostic grading of DR according to ETDRS guidelines.

Wang et al. showed in the “Retinopathy Online Challenge” that their approach detected microaneurysms better than other approaches, with an average score of 0.464 (Wang et al., 2017a). However, there seems to be room for improvement, considering that mild DR, for example, presents with microaneurysms only in an otherwise healthy retina. The distinction between mild and no DR is also the hardest for a clinician as single microaneurysms may be easily missed. Other clinically more distinctly visible structures show superior results for automated feature detection. The sensitivity and specificity of detecting a hemorrhage was 79% and rose to 92% when the task was to identify images where hemorrhage was present (van Grinsven et al., 2016). This compares to the clinician whose task would be to state “yes, there is hemorrhage” as opposed to “there is hemorrhage in the superior upper quadrant of the retina,” which is certainly a more difficult task. This example highlights the fact – clinically and automated – that it is in general easier for a clinician/algorithm to make a correct decision on an entire image than solely on presentation of an isolated feature at a certain position. It emphasizes that algorithms benefit from “context,” as do ophthalmologists. This can also be confirmed for hard exudates, where the sensitivity was 92% for individual feature detection and 100% for correct classification of the entire image (García et al., 2009b).

Obviously, there are other relevant applications besides DR for the use of automated feature detection in retinal images. The cup to disc (c:d) ratio is the most important feature for use in glaucoma detection (Fan et al., 2018; Haleem et al., 2016; Haleem et al., 2017; Miri et al., 2015). As the clinician usually looks at the c:d ratio to judge the risk or progression of glaucoma, algorithms can be used for the same task. Fan et al. could show an accuracy of 98% in the detection of the entire optic disc, which is the most important first step before the c:d can be measured (Fan et al., 2018). These investigators applied a multimodal approach to detect the disc and cup simultaneously in CFP and OCT with solid results, but the system has only been applied to 25 cases and is therefore not sufficient for providing a valid AI method clinically, which usually requires big datasets (Miri et al., 2015).

In general, numbers for sensitivity and specificity have to be weighed carefully in AI diagnosis. The detection of single features can be less accurate. Nevertheless, the entire image can be judged with diagnostic accuracy and the correct treatment decision made. For comparison, a single microaneurysm during slit-lamp examination may be missed, but this will most likely not have any clinically relevant consequences. However, in severe DR with many pathological features present simultaneously, the examiner – an ophthalmologist or an algorithm – may miss an individual microaneurysm and only detect 9 out of 10 and the correct diagnosis of severe disease will still be confirmed.

3.1.2. Detection and Quantification of features

Automated detection of CNV, a fibrotic scar, atrophy or drusen in CFP are relevant in the context of clinical classification in AMD. These features can either be solely detected or quantified, which means measuring areas and volumes, e.g., of drusen (van Grinsven et al., 2013), pseudodrusen (van Grinsven et al., 2015) or geographic atrophy (Feeny et al., 2015). Van Grinsven et al. quantified drusen in 407 images of patients with AMD specifying the location, area and size of each druse. The main focus was an automated AMD risk assessment predefined by a central reading center (van Grinsven et al., 2013). This assessment achieved an accuracy of 95%, similar to the performance of two independent human graders. The drusen area between the two human observers achieved an agreement of κ=0.87 and the algorithm compared with each human observer reached a κ=0.91 and κ=0.86. The same group tested an automated algorithm for the detection of reticular pseudodrusen in a multimodal approach with CFP, fundus autofluorescence and near-infrared fundus images in 230 cases and achieved an accuracy of 94% (area under the ROC curve) (van Grinsven et al., 2015). This output clearly supports the notion that machine learning approaches can well be used for risk determination in AMD at the level of human specialists. Moreover, novel insight into the pathophysiology of AMD was established implying that drusen volume regression can be a predictor of late AMD stages. The feasibility of such efforts confirms again the enormous potential of such automated approaches, as no human expert would be able to track the drusen volume over time manually.

In OCT, the detection of IRC and SRF (Figure 16) is most relevant when considering exudative diseases such as AMD or DME (Schmidt-Erfurth et al., 2018a; Schlegl et al., 2018b). Deep learning has been applied for a binary decision whether disease activity is present or not – as necessary for the clinician's decision to "treat or not to treat" in a flexible, PRN or treat-and-extend (T&E) regimen. IRC or SRF may be important when determining disease activity from OCT scans or for initial diagnosis. Algorithms which detect whether fluid is present in AMD or DME can serve for diagnostic grading of disease activity at baseline (Chakravarthy et al., 2016; Liu et al., 2011a; Sidibé et al., 2017; Srinivasan et al., 2014; Sun et al., 2017) to facilitate clinical processes (e.g. preselection of the most important scan for the examining doctor to save time (Chakravarthy et al., 2016) and for treatment decisions (Schmidt-Erfurth et al., 2018a; Prahs et al., 2018). Please refer to the chapter "Guidance of therapy" for further details.

Alsaih et al. compared the different algorithms available for DME detection in OCT and found a very variable sensitivity/specificity in differently sized datasets (69%/94% in 45 (Srinivasan et al., 2014), 81%/63% in 32 (Sidibé et al., 2017) and 69%/94% in 326 (Liu et al., 2011a) compared with their own results of 88%/88% in only 32 datasets of 16 patients and 16 healthy controls (Alsaih et al., 2017). Efficiency in a small number of patients may be achieved, however, a high accuracy for adequately training the algorithm can usually only be reached in large datasets. Again, this is similar to a human progressively gaining expertise, e.g., in distinguishing DME or neovascular AMD during residency. A retina specialist usually and "intuitively" will not have any difficulty with the correct diagnosis of an OCT based on prior experience with large numbers of patients.

Sun et al. evaluated images of DME and dry AMD compared with healthy controls and automatically classified 99.7% of the images correctly as diseased or healthy. Their dataset consisted of 297 DME scans, 213 healthy scans and 168 dry AMD OCT scans (Sun et al., 2017). This result is superior to the results of the risk assessment of van Grinsven et al. in terms of correct detection but when considering its value in clinical routine most likely less relevant than the risk-specified assessment (van Grinsven et al., 2013). A third group classified AMD with drusen detection from OCT and CFP scans combined and were most successful in achieving a sensitivity/specificity of 100%/97% in 100 CFP and 6,800 OCT images of 100 patients (Khalid et al., 2017). This underlines the importance of an algorithm having adequate practice for clinical usage and the importance of a solid learning part with a large training set for any machine learning model.

With regards to detection and in addition grading stages of geographic atrophy, results using AI in OCT instead of CFP are already most convincing. Ji et al. evaluated an algorithm against two manual expert graders and achieved a κ>0.99 for each comparison (human 1 vs. human 2, human 1 vs. algorithm, human 2 vs. algorithm), which offers promising clinical applicability for the support of clinicians in daily routine (Ji et al., 2018). The distinction between individual retinal layers is also of relevance in this context (ElTanboly et al., 2017). An alteration of the RPE, as in PED, may help to guide the diagnosis in a screening situation for active AMD (Sun et al., 2016). Nerve fiber layer changes are important for glaucoma detection by OCT, where the c:d ratio can also be detected in great morphological detail (Miri et al., 2017).

In conclusion, the quantification of features, for example of fluid, is more informative than a simple binary yes-no classifier in diagnostic grading of exudative disease. Different groups have used AI-based algorithms, mostly supervised machine learning approaches, to manage this difficult task particularly for fluid quantification of IRC or SRF volume (Figure 16) (Breger et al., 2017; de Sisternes et al., 2017; Lee et al., 2017b; Montuoro et al., 2017; Roy et al., 2017). Another important feature that may benefit from quantification and has been solved with AI approaches is the volume of drusen in early to intermediate AMD, as this appears to be an indicator for the risk of disease progression (Abdelfattah et al., 2016; Chen et al., 2013; Bogunovic et al., 2017a; Schmidt-Erfurth et al., 2018b). Another promising application of AI in grading disease severity may be the detection and measurement of subclinical features such as hyperreflective foci (HRF) that are only visualized by OCT resolution and not opthalmoscopically. Quantification of HRF is relevant for both AMD and DR grading. Although both have different pathophysiological origins such as migrating RPE or lipid exudation, they can be detected with the same algorithm and support disease classification or risk assessment of different retinal diseases more reliably than the human expert can ever achieve.

3.2. Screening for Retinal Disease

The aim of screening is to differentiate subjectively healthy individuals into (many) objectively healthy people and (few) objectively diseased patients. Screening is a preventive method and useful if it achieves high sensitivity and specificity and if the output of the screening purpose proceeds in a meaningful context. This may, for instance, be a shortening of the delay to detection in an elevated risk for a vision-threatening event such as CNV in a fellow AMD eye or the timing of a therapeutic intervention. Screening is primarily the binary decision between healthy and diseased (the distinction of different activity stages will be described in the following chapter). In terms of setting, screening is most efficient in large groups of individuals who would not be seen by an ophthalmologist in regular practice. Screening technology could become widely accessible with the era of smartphones. Dedicated screening cameras will most likely be integrated into portable digital devices. Such procedures should be autonomous with little requirement for trained personnel and should ideally be affordable and time-saving for the “incompliant” patient who fails to make regular visits to a doctor’s office.

The first step necessary in a well-designed screening algorithm is recognition and exclusion of insufficient image quality. Poor quality could lead to a false negative result and may prevent timely diagnosis. In addition, an algorithm’s sensitivity and specificity are the most important determinants for deciding whether it can be used for screening purposes. A high sensitivity is important to avoid missing patients who urgently require treatment as they were false negatives. A high specificity helps to prevent the healthcare provider from being flooded with too many false positives and undertaking inefficient interventions. False decisions could also increase patients’ and doctors’ mistrust in the usefulness of a screening tool. Screening devices should ideally be located in places with high patient through-put such as primary care offices, which are not routinely involved in ophthalmological care but could support ophthalmic screening when equipped with reliable AI tools.

3.2.1. Diabetic Retinopathy Screening

Early detection of retinopathy is an important part of management for millions of individuals with diabetes. According to the International Diabetes Federation there were 425 million people with diabetes worldwide in 2017. The current guidelines of the American Diabetes Association recommend patients with diabetes without any eye symptoms see an ophthalmologist bi-annually. However, about 50% of these patients do not follow the recommendation. The most common and most needed use case for retinal screening with AI methods is thus the screening for DR. Several research teams and companies therefore focus on DR screening from CFP (Abramoff et al., 2013; Abramoff et al., 2016; Gargeya and Leng, 2017; Gulshan et al., 2016; Ting et al., 2017) (Figure 17).

Screening for DR means the classification of patients with diabetes as an underlying disease into patients with no DR (with no retinal changes) and patients with DR, where few or many retinal changes can be detected even in the absence of any visual complaint. Several clinical devices are currently under evaluation or already available (for instance, RetinaLyze®). At the moment, these are still large fundus cameras, designed for use in the field of ophthalmology and not yet small portable smartphone additions. In April 2018, the US FDA permitted marketing of the first medical device to use AI to detect greater than a mild level of DR (iDx-DR). The iDx-DR device combines an AI- and cloud-based algorithm with an almost autonomous retinal fundus camera. Fundus images of sufficient quality are automatically differentiated into negative (= non-referable = no or mild DR) or positive, indicative of a condition of more than mild DR resulting in the referral to an ophthalmologist (referable DR). Their algorithm does not differentiate between no or mild DR as the presence of a few microaneurysms (mild DR) would not lead to any clinical consequences for the patient (Abramoff et al., 2013; Abramoff et al., 2016). FDA approval was granted based on a study of 900 patients with diabetes at ten primary care sites, which resulted in correct identification of a positive finding indicative of DR in 87.4% of individuals and a correct negative result in 89.5%. As the detection algorithm was trained on DR in untreated otherwise ophthalmologically healthy patients, previous laser and surgical or pharmacological treatment were exclusion criteria as well as manifest disease with DME, severe non-proliferative or proliferative DR. Furthermore, comprehensive eye examinations were recommended at the age of 40 and 60 years. Interestingly, the FDA set a mandatory level of accuracy as the primary endpoint for this trial with a sensitivity of more than 85% and a specificity of more than 82.5%. Considering this context, the accuracy numbers given on different screening methods may be interpreted regarding the IDx device as a benchmark. In comparison, particularly the sensitivity of ophthalmoscopy performed by clinicians is substantially lower at 73%, with a 91% specificity (Lawrence, 2004). Thus, we may interpret that the FDA chooses a higher sensitivity over accurate specificity in the context of screening applications.

In another published DR screening algorithms, the sensitivity varied from 87%-97%, and the specificity from 59%-98% (97%/59% (Abramoff et al., 2013), use of 1748 images; 97%/87% (Abramoff et al., 2016), use of 1748 images; 87% (Gulshan et al., 2016), use of 1748 images; 91-100%/91%-92% (Ting et al., 2017), use of 76,370 images; 94%/98% (Gargeya and Leng, 2017), use of 75,137 images). These numbers show that the majority of available AI methods would be capable of being used for DR screening according to requested FDA endpoints and most of them seem to be performing better and faster than clinicians.

The largest study to date among the publications mentioned was performed by Ting et al., who reported the development and validation of a deep learning system for DR detection using CFP images from the Singapore National Diabetic Retinopathy Screening Program. The system was evaluated in 494,661 images for the primary disease DR and also for concomitant diseases including glaucoma and AMD. The sensitivity and specificity for referable DR, vision-threatening DR, glaucoma and AMD were all above the FDA criteria (91%/92%, 100%/91%, 96%/87%, 93%/89%). In the same study, 10 multiethnic datasets were tested for referable DR as well and here the sensitivity ranged from 92%-100% and the specificity from 73%-92%. The ground truth was developed within national screening program gradings by human graders (Ting et al., 2017).

All these examples show that AI-based DR screening algorithms have reached or may even outperform the level of accuracy of clinical experts. DR screening in particular carries enormous potential as support for ophthalmologists, may help to reduce the prevalence of late and cost-intensive disease stages and is likely to pioneer digital medicine applications in the near future and at a large scale.

3.2.2. Screening for Age-Related Macular Degeneration

Another retinal disease with pandemic dimensions that lends itself to AI-based screening is AMD. Several publications report on the automated detection of AMD on color photographs with a sensitivity/specificity of 93%/89%, use of 72,610 images, or an accuracy of 90% (“comparable to human performance”), use of more than 130,000 images (Burlina et al., 2017). Approaches based on fundus photographs may be clinically useful in the diagnosis of early AMD stages to identify patients requiring more detailed investigations and follow-up. Nevertheless, compared with DR screening, the role of color photography in the diagnosis of AMD, particularly neovascular AMD, is rather limited and it is reasonable to believe that OCT-based screening methods could be more successful in this specific setting. OCT is able to identify several signs of AMD that are not visible on fundus photographs, including hyperreflective foci and outer retinal thinning, which is an early hallmark of geographic atrophy. Detecting these signs may be helpful in identifying patients with AMD disease as opposed to individuals presenting normal retinal ageing. However, an important drawback of OCT, at least at this time, is its limited availability in the primary care/screening setting. The introduction of low-cost OCT systems will probably allow more screening tools to be available based on this technology.

AI research using OCT to screen for AMD includes that of Venhuizen et al. They published a screening system validated in 367 individuals which reached a sensitivity and specificity each of 93% against their reference standard (examination by an ophthalmologist). This screening system was developed on a fairly large database of 3,265 OCT scans (Venhuizen et al., 2017).

3.2.3. Glaucoma Screening

Algorithms use the same features clinicians evaluate during slit-lamp examination for glaucoma screening. Abnormalities of three kinds are used: the c to d ratio, the area of the neuroretinal rim and the "ISNT" rule, evaluating the width distribution of the neural rim around the optic nerve head (Haleem et al., 2017; Issac et al., 2015; Kim et al., 2017).

Isaac et al. reported on an AI-based glaucoma classification using all three features for the distinction between healthy individuals and patients with glaucoma. They achieved an accuracy of 94% as well as a sensitivity and specificity of 94% in detecting glaucoma. As a limitation, this was only tested in a dataset of 67 eyes (35 healthy, 32 glaucoma). In clinical practice, perimetry and OCT measurements are further contributors to an early diagnosis of glaucoma. The combination of the above-mentioned three features with changes in the retinal nerve fiber layer has large potential for screening purposes as well but the same limitation mentioned as for AMD (i.e. the limited accessibility of OCT) must be considered. Slit-lamp features together with visual field analyses improve results of glaucoma screening to a large extent: in a dataset of 499 individuals (297 glaucoma, 202 healthy; 98% accuracy, sensitivity and specificity) results were excellent but the practical use for screening purposes is limited as visual field data are not available for population-based screening programs and are certainly not the modality of choice when considering low-cost broadly available and easy-to-use devices (Kim et al., 2017).

Bowd et al. also applied their algorithmic approach to the combination of OCT and visual field data in a dataset of 69 healthy control participants and 156 patients with glaucoma and tested their algorithm for OCT alone, visual field alone and a combination of both (Bowd et al., 2008). This led to an accuracy of 82% (OCT alone), 84% (visual field alone) and 87% (combination) (area under the ROC curcve) and showed that there is potential for stand-alone OCT approaches for glaucoma screening as well as for CFP. In CFP images, the study by Ting et al. mentioned earlier detected possible glaucoma as a concomitant disease in a huge dataset of patients with diabetes with an accuracy of 94% (area under the ROC curve; use of 125,189 images (Ting et al., 2017).

3.2.4. Retinopathy of Prematurity Screening

Similar to DR, retinopathy of prematurity (ROP) is a condition that can be well-diagnosed from CFP alone. Human-graded screening programs or telemedicine are widely used in less specialized hospitals. The largest established ROP screening program is the SUNDROP (Stanford University Network for Diagnosis of Retinopathy of Prematurity) program established in 2008 by Moshfeghi et al. and followed up in publications for 6 years (Murakami et al., 2008; Wang et al., 2015). However, ROP screening has not yet seen large developments regarding automated, AI-based diagnosis.

There have been a few attempts at using machine learning to automatize ROP diagnosis. Ataer-Cansizoglu et al. developed the i-ROP system (Ataer-Cansizoglu et al., 2015). Here, AI was applied on a dataset of 77 wide-angle retinal images of infants and compared with clinical judgement in ophthalmoscopic examinations as well as three manual gradings of each image. The i-ROP system achieved 95% accuracy for classifying ROP versus no ROP compared with fundus examination. This was comparable to the performance of three individual experts (96%, 94%, 92% accuracy). In a second approach the i-ROP system was evaluated against more experts with an accuracy of 79%-99% when fewer features where used for diagnosis. The original accuracy of 95% could not be achieved with less features, reassuring that the use of all known features results in the most robust ROP diagnosis (Campbell et al., 2016). It needs to be mentioned that the gold standard for the detection of ROP for clinical application is the examination by a ROP specialist and CFP will already be inferior because important details of the retinal periphery may be missed. Both of the studies mentioned have only compared with the specialist grading of the image, not with clinical examination. Again, the number of available clinical experts in unspecialized clinics (that may be specialized for survival of prematurely born infants but may not have the availability of a retina specialist) has to be compared with the less accurate CFP screening and for manual human-graded programs this deficit has been accepted. Therefore, there is a very large potential for the establishment of automated screening programs for ROP in the future as most clinics have cameras already available and, as in DR, an accurate result can be achieved for these images. Nevertheless, it may be harder to establish such programs as the accuracy is usually set higher in medical needs and symptoms that cannot be clearly articulated and where a false-negative results can be much more devastating than in DR.

The use of ROP screening will remain a relevant future topic as there are many medical centers that may be able to deliver general healthcare to prematurely born infants, but may not have an ophthalmic expert available for continuous monitoring, especially early after birth when infants are still hospitalized. Here automated or human-graded will be a decision of cost effectiveness and need.

3.2.5. Screening for retinal disease in general

Many groups currently work on the development of image-based screening for retinal disease both on CF and OCT and the future will certainly lead to such approaches for quick filtering of healthy persons and those with disease. This will allow the ophthalmologist to concentrate on the management of clinically affected patients, monitor them for high risk of disease development and initiate therapy in a timely manner. Approaches for the classification of CFP and OCT images for no retinal disease versus retinal disease that can serve for screening purposes without limitation to a specific alteration have already been published (Choi et al., 2017; Liu et al., 2011b).

Choi et al. described automated differentiation between normal eyes and nine retinal diseases from CFP images alone (Choi et al., 2017). The study lacked a large database (only 397 cases of which only 25 cases were healthy) and the cases were unevenly distributed among diseases (one disease had only one image while another had over 60 images). This is most likely the reason for only 31% accuracy for all ten classes. When only the three most common classes were classified (AMD, DR and normal), an accuracy of 73% could be achieved. Using large real-life datasets from hospitals, for example, will most likely allow a much higher accuracy in the near future.

Liu et al. described a screening approach in OCT images for differentiation between healthy eyes and eyes with macular edema, macular hole or AMD. 193 eyes were used for development and 131 for validation of classification into these four categories. The accuracy (area under the ROC curve) was between 93% and 98%, depending on disease (highest for healthy cases) which is a promising result (Liu et al., 2011b).

3.3. Diagnostic grading/staging of retinal disease

The difference between screening and diagnostic grading or staging of retinal disease mainly lies in the patient population. Screening is applied in subjectively healthy individuals, is performed in large cohorts and will usually have only a small proportion of positive test results throughout the population. DR screening is performed to distinguish between people with diabetes and no signs of DR and people with diabetes and signs of DR. Consecutively, people with signs of DR can be divided into people with non-referable signs of DR (these do not need to see an ophthalmologist and have mild DR) and patients with referable DR. This positive subpopulation can be further divided resulting in diagnostic grading or staging. Patients with referable DR can have moderate DR, proliferative DR or DME. Further, they can have vitreomacular traction, an epiretinal membrane or retinal detachment. This high potential of differentiation clearly highlights the enormous savings in time and manpower introduced in a large scale of health care monitoring, which a developed society should be able to grant all of its citizens. These are all summarized in the following chapter discussing diagnostic grading/staging for retinal disease in the different diseases beyond screening. Screening and diagnostic grading/staging may happen practically at the same time of presentation but the development of diagnostic grading/staging algorithms requires many more datasets. For example, a study for screening may require 3,000 sets to have a fair amount of disease in the sample (e.g. 5% = 150 cases). But a division of 150 cases into three different classes only results in 50 samples in each group. Considering that for validation half of the samples are used as the test set and the other half for validation, it is obvious that a screening algorithm could be developed faster with less patients.

3.3.1. Diagnostic grading/staging of DR

As mentioned in the chapter *Diabetic Retinopathy Screening*, Abramoff et al. distinguish between referable and non-referable DR as do most automated DR algorithms. A staging between moderate DR and vision-threatening DR is performed for their referable DR cases. Vision-threatening DR includes severe non-proliferative DR, proliferative DR and DME according to the ICDR classification (International Clinical Classification System for Diabetic Retinopathy by the American Academy of Ophthalmology). The further automated differentiation between these cases is not delivered as an output by the device as there is no practical consequence and a false positive result may result in wrong assumptions for a patient with clinical and legal implications. Therefore, the ophthalmologist must serve as an expert for therapeutic consequences and again the screening output should ideally only support the clinician and not entirely take over the therapeutic consequences (Abramoff et al., 2016). The legal aspect, however, is an issue which has not been clarified yet, as doctors may be sued for errors and algorithms necessary need legal control as well.

Gulshan et al. tried diagnostic grading/staging on their two datasets (8788 images and 1745 images) as well. For "moderate or worse DR only" the sensitivity was 90% and 87%, respectively, and the specificity was 98% in both datasets; for "severe or worse DR only" the sensitivity was 84% and 88% and specificity 99% and 98%; for "DME only" the sensitivity was 91% and 90% and the specificity 98% and 99%. This shows that having a large dataset allows acceptable results for advanced disease stages as well (Gulshan et al., 2016), as Ting et al. confirmed using almost half a million images with a sensitivity and specificity for referable DR versus vision-threatening DR only (excluding moderate DR) of 91%/92% versus 100%/91% (Ting et al., 2017).

Takahashi et al. focused on diagnostic grading/staging by using the ground truth of actual interventions (laser, injections, surgery, nothing) performed after an image was taken (Takahashi et al., 2017). They included 4709 CFP and categorical visual acuity changes (improved, stable, worsened) for training and tested the algorithm on 496 cases, reaching an accuracy of 96% in the prediction of interventions compared with three retina specialists who reached an accuracy of 92-93%. This is a very practical concept and can be relevant in making adequate treatment decisions. Nevertheless, the false-negative rate – when the grade was "not requiring treatment" but treatment was actually needed – was 12%. The false positive rate – when the grade was "requiring treatment in the current visit" but treatment was actually not needed at the next visit – was 65%, which can lead to a large number of visits for treatment which might not be needed. This is not only cost ineffective but also creates a large number of alarmed patients who believe a treatment will be needed.

3.3.2. Diagnostic grading/staging of AMD

Diagnostic grading/staging is also an important AI application in AMD as the changes seen in OCT determine the progressive stage of disease and no binary cut-off can be made. Changes seen in early AMD can remain for decades without progression. Venhuizen et al. analyzed AMD OCT data for screening purposes, not only in a binary approach as mentioned earlier but also dividing 367 individuals into 5 different grades for diagnostic grading/staging: no AMD, early AMD, intermediate AMD, advanced AMD geographic atrophy and advanced AMD choroidal neovascularization (Figure 18). The overall sensitivity and specificity reached 98% and 91% against the reference standard (examination by an ophthalmologist). Depending on the different diagnostic stages, different treatments and individual prognoses will be the consequence which requires advanced medical and legal control (Venhuizen et al., 2017). Another way of diagnostic grading in AMD is drusen phenotyping, which can serve as the basis for prediction and risk assessment of disease conversion as outlined in the chapter on prognosis.

3.3.3. Diagnostic grading/staging of Glaucoma and Retinopathy of Prematurity

No system to differentiate between different stages in glaucoma has been published. Publications in glaucoma "staging" focus on the progression of disease as this indicates that (more) therapy is necessary, even if the intraocular pressure might be in a normal range. Potential can be seen in monitoring patients over time and taking changes of the c:d ratio and the nerve fiber layer thickness into consideration for learning models. To date, no such study in a large patient cohort using machine learning has been reported.

The few available publications for the use of machine learning in ROP have focused on the differentiation between no ROP and (pre-)plus disease. The different zones/vascularization stages have not been the focus of any large patient cohort using machine learning in current publications. The potential would be in automatically classifying these different stages of vascularization for monitoring purposes and timely treatment indication in a setting lacking regular expert surveillance. This would be a reliable way to trigger treatment decisions – similar to AMD, DME or retinal vein occlusion – especially in anti-VEGF treatment in premature infants (once entirely established). Ataer-Cansizoglu et al. distinguished between no ROP, preplus disease and plus disease but did not take the different vascularization stages into account (Ataer-Cansizoglu et al., 2015; Campbell et al., 2016).

3.3.4. Diagnostic grading for systemic disease

As has been recognized early on, the retinal condition may reflect systemic disease. Many ophthalmologists are routinely involved in diagnosing systemic diseases such as hypertension, sarcoidosis and other autoimmune diseases, syphilis, CMV infection, tuberculosis etc. from ophthalmoscopy. Poplin et al. demonstrated in an impressive way how many and various conditions can be recognized using AI in fundus images. The diagnostic grading of retinal CFP images to search for cardiovascular disease was trained on images of 284,335 patients with the primary objective of predicting cardiovascular risk factors. The outcomes showed moderate accuracy in the primary objective but other features were identified with high accuracy: age (accuracy ±3.26 years), sex (accuracy 97%), smoking status (accuracy 71%), systolic blood pressure (mean absolute error within 11 mmHg) and major cardiac adverse events (accuracy 70%) (Poplin et al., 2018). This list of features demonstrates that diagnostic grading from retinal images reaches far beyond retinal disease and that automated algorithms will enable us to detect more from images than any clinician would be able or intend to diagnose. AI analysis can detect subclinical and discrete features appearing below the threshold of a human observer, quantify minimal differences in feature expression and recognize patterns among large cohorts. When broadly available, future indications will likely include vascular pathologies, ageing disorders and neurodegenerative diseases such as Alzheimer's disease and multiple sclerosis, not only in CFP, but also in OCT (e.g. OCTiMS study: Optical Coherence Tomography (OCT) Trial in Multiple Sclerosis, a 3-year, pharmacologically non-interventional study to evaluate OCT as an outcome measure in patients with relapsing remitting multiple sclerosis; clinicaltrials.gov study identifier NCT02907281). The retina as a window to the body will attract a lot of non-ophthalmological attention, once "exploited" extensively by AI methodologies.

3.4. Guidance of therapy

One major advantage of AI, particularly for designing an optimal therapeutic management, is that it enables individual clinicians to access and utilize prior experience provided by hundreds of thousands of previous cases. AI generates knowledge from data in a much more accessible and reproducible way than the most experienced experts. By detecting characteristic patterns in large datasets, machine learning as opposed to population level studies, for instance, offers ground-breaking progress in the field of personalized prognosis. Analogous to precision medicine, where, for instance, oncologic therapy is prescribed according to the specific individual tumor genotype, AI in ophthalmology may enable an individualized prognosis of therapeutic response, optimal retreatment intervals, and future disease progression. Furthermore, AI could be used in a more traditional sense to automatically diagnose disease activity in retinal imaging data to blankly automatize and standardize office procedures and retreatment assessments. Thus, a standard of best practice could be reliably implemented in any setting in a cost and time effective manner. Finally, advanced disease models based on AI are able to provide valuable insight into the pathophysiology of disease by interpreting the microstructural features used by predictive analyses.

3.4.1. Automated detection of disease activity

Growing patient populations in times of progressive longevity in industrialized countries, the increasing prevalence of retinal disease and widening of the retinal therapeutic spectrum continue to challenge the daily practice of ophthalmologists with an overwhelming number of visits and imaging investigations. In this setting, it is demanding for clinicians to consistently assess the large number of images per patient in a reliable and time-efficient manner. Furthermore, disagreement exists over adequate interpretations of the changes seen in retinal imaging studies. For instance, reading centers in a trial setting may grade OCT images differently from study investigators or, even more, physicians in the real world (Heimes et al., 2016; Toth et al., 2015). Here, AI may provide urgently needed relief by providing automated, standardized assessment of disease activity to improve clinical management and usage of healthcare budgets.

Deep learning to diagnose disease activity in OCT images of patients with neovascular AMD was proposed by Chakravarthy et al. in 2016 (Chakravarthy et al., 2016). The algorithm presented detects overall presence or absence of macular fluid in Cirrus OCT images with an accuracy reaching close to the inter-observer agreement of three retina specialists or a reading center. Furthermore, the algorithm has been used to highlight the OCT slices containing the most relevant information regarding the presence of fluid, which was suggested to enhance the capability of the ophthalmologist to focus on these scans in a time-effective manner.

Recently, Prahs et al. proposed a similar deep learning model, however, with the goal of automatically determining the need of anti-VEGF retreatment rather than purely the presence of fluid (Prahs et al., 2018). The model was trained on over 180,000 central B-scans of patients under real-world anti-VEGF therapy and corresponding retreatment decisions. A predictive performance of over 95% was achieved. A major drawback of this work is that only central B-scans were considered, which may provide false results in the case of scan misplacement, e.g., during inability to fixate centrally and neglect justafoveal pathologies. Similarly, approaches to automatically detect disease activity in DME have been presented but based on very limited validation in a small dataset (Alsaih et al., 2017).

In general, despite the excellent performance of AI algorithms in classifying disease activity, the main limitation of these approaches is that they only provide a binary decision regarding the presence or absence of fluid. This lack of granularity may hinder useful application in clinical practice, as outlined below.

3.4.2. Automated quantification of pathology

In the management of patients with retinal disease, it is mostly not the presence or absence but the quantity of a particular pathology that determines therapeutic decision making. For instance, in diabetic macular oedema, current treatment recommendations include to administer anti-VEGF therapy until fluid remains stable (Wells et al., 2015). Another example is the recommendation to treat PED if it exhibits active growth which is reflected in an increase in PED volume (Penha et al., 2013; Schmidt-Erfurth et al., 2015). Furthermore, investigators suggest differential roles for different types of fluid (Schmidt-Erfurth and Waldstein, 2016). For instance, intraretinal fluid may be a retreatment indication, while subretinal fluid up to a certain threshold (200 micrometers in height at the foveal centre) may not (Arnold et al., 2016). All these paradigms require some degree of differentiation and quantification of the microstructural changes in the retina. Moreover, quantification of pathology could be important for prognostic reasons as some biomarkers (such as intraretinal fluid) show a tight correlation with visual acuity and vision outcomes (Waldstein et al., 2016). In intermediate AMD, measurements of drusen volume could be used to assess the individual risk of CNV onset (Abdelfattah et al., 2016; Schmidt-Erfurth et al., 2018b).

Several groups have addressed this unmet need and have started to develop algorithms to automatically quantify retinal pathology using AI. One important target of quantification is retinal fluid and a few successful approaches have been presented, mainly based on supervised deep learning (Breger et al., 2017; Lee et al., 2017b; Montuoro et al., 2017; Roy et al., 2017). The method proposed by Schlegl et al. is the most extensively validated and most broadly applicable (1,200 eyes, 3 diseases, 3 OCT machines), achieving an overall accuracy of R2=0.90-0.96 (Schlegl et al., 2018b). It also differentiates between subtypes of fluid, which has important implications on prognosis and management (Figure 16). Other approaches often suffer from limited validation in relying on only a few cases or a narrower selection of diseases or devices.

A second relevant setting for the quantification of disease is dry AMD in both its early and late forms. Using AI, it is possible to quantify drusen on OCT (Chen et al., 2013; de Sisternes et al., 2017), yielding drusen volume measurements rather than drusen area alone when alternatively based on fundus photography (Rubin et al., 2013; van Grinsven et al., 2013). However, the quantification of pseudodrusen currently seems to be confined to 2D imaging methods (van Grinsven et al., 2015).

Investigators have also proposed quantification of hyperreflective foci using deep learning (Schlegl et al., 2018a). Large foci may correspond to pigmentary changes on color photographs, when quantified using deep learning (Schmitz-Valckenberg et al., 2016). HRF were shown to represent a major biomarker in dry AMD progression (Schmidt-Erfurth et al., 2018b). In this respect, AI methods nicely reflect histological evidence showing RPE migration in active disease (Curcio et al., 2017). Moreover, segmentation algorithms have been developed for quantification of GA lesions on 3D OCT as well as on fundus photographs (Feeny et al., 2015; Ji et al., 2018). It has become clear that fundus autofluorescence (FAF) merely demonstrates end stage findings with a black RPE defect seen in 2D-en-face FAF images, while 3D SD-OCT depicts primary neurosensory loss together with RPE migration preceding active GA progression (Sayegh et al., 2017).

3.4.3. Prediction of need for retreatment

One of the great dilemmas of intravitreal therapy is the difficulty in determining and planning adequate retreatment intervals. In an ideal world, patients would receive intravitreal treatment as often as required to maintain complete disease control but as rarely as possible to avoid the potential morbidity associated with anti-VEGF therapy such as endophthalmitis or development of RPE atrophy. To achieve this goal, several therapeutic regimens have been proposed in the community, including pro-re-nata (PRN) as well as treat and extend (T&E). Both regimen have resulted in non-inferior visual acuity outcomes compared with monthly therapy in randomized controlled trials (Busbee et al., 2013; Silva et al., 2018). However, in clinical practice, a PRN approach results in monthly follow-up visits, virtually for a lifetime, which are clearly not manageable for both patients and physicians. Despite the frequent visits, slow but irreversible visual decline may result from the frequent recurrences that are essentially required to retreat the patient (Schmidt-Erfurth et al., 2015) as well as out of potential delays between the diagnosis of an exudative recurrence and subsequent retreatment (Ziemssen et al., 2016). On the other hand, T&E schemes offer pragmatic scheduling, a reduced number of visits and avoidance of long treatment-free intervals. Nevertheless, concerns include the potential overtreatment of patients who have biologically low requirements for anti-VEGF and problems associated with trying to extend individuals who are per se non-extendable and who are exposed to the risk of exudative events by protocol (Freund et al., 2015).

In this scenario, the use of AI methods promises to result in predictive models that can determine upfront the need for treatment requirements and frequencies. Ideally, an AI model would take the images and clinical characteristics of a given patient acquired at baseline as well as after a first injection and would provide, e.g., the probability for extendibility up to a certain interval, i.e., the optimal extension length as well as overall expected therapeutic requirements over a certain time frame. Implemented in clinical practice, such models could dramatically improve the plannability of anti-VEGF therapy, including healthcare expenditure control and help to appropriately manage expectations of patients and physicians, and finally result in better outcomes due to avoidance of under- or overtreatment.

The individualized prediction of optimal, personalized retreatment intervals has already been achieved in proof-of-principle studies in the field of neovascular AMD. Recently, Bogunovic et al introduced an AI model based on random forest that was designed to classify patients with a low, medium or high need for retreatment a priori (Figure 7) (Bogunovic et al., 2017b). Data of 317 patients receiving ranibizumab PRN therapy in a randomized, controlled multicenter trial were included for model training. The OCT images acquired during the common loading dose (month 0-3) were analyzed using image analysis algorithms based on deep learning and graph cut (Schlegl et al., 2018b; Zhang et al., 2014). This resulted in several hundred quantitative, spatially and temporally resolved variables describing the individual retinal morphology and the initial therapeutic response of each patient. The resulting features were introduced into a modelling database and used for machine learning. Over the remaining 21 months of the trial, 22% of patients showed a low need for retreatment (0-6 injections), 56% medium treatment requirements (6-15 intravitreal injections) and the remaining 22% exceptionally high retreatment needs (16-21 injections). With the AI model, it was feasible to a priori differentiate these three groups with an accuracy (AUC) of 70%-77%.

Noteworthily, the performance of the automated algorithm was by 50% more accurate than the assessment of a human retina specialist, particularly in determining patients with a high therapeutic need in the future, hence even in predictive challenges AI methods outperform experts.

Moreover, a view into the ranking of the clinical relevance of input features in the random forest model offers an unbiased insight into the most relevant OCT biomarkers determining overall retreatment need. Specifically, the amount of subretinal fluid remaining at the end of the loading dose ranked highest and high volumes of SRF were significantly associated with a future need for more frequent retreatments. However, for post hoc analyses like this, AI can only reproduce the intentions of the protocol which required mandatory injection in any type of fluid. One cannot conclude that SRF resolution is mandatory for visual recovery.

A similar predictive tool based on random forest and convolutional neural networks was recently developed for the prediction of T&E intervals in the therapy of patients with neovascular AMD (Bogunovic et al., 2018). Patient-level data of 210 eyes receiving ranibizumab according to a T&E regimen or at 12-month intervals were used for this study. The AI model received automatically determined, quantitative OCT biomarkers at baseline and after the first injection for training. The goal of prediction was to classify extendable (injection interval between 8 and 12 weeks, 82% of the cohort) versus non-extendable patients (interval between 4 and 6 weeks, 18% of cohort). Furthermore, the investigators attempted to predict the maximum fluid-free interval during the course of the trial. The model was successful in determining extendable versus non-extendable patients with an accuracy (AUC) of 75%. The prediction of the longest recurrence-free interval was more challenging at R2=0.27. Similar to previous findings, the volume of subretinal fluid remaining after a first treatment represented the most important biomarker considered by the model based on the definitions of retreatment by protocol.

Investigators have also proposed predictive models for future retreatment in the field of macular edema secondary to RVO. Vogl et al. reported on an AI-based predictive tool that could determine future recurrence of macular edema after the loading dose with an accuracy of 79%-83%, based on 247 eyes with a 12-monthly standardized follow-up (Vogl et al., 2017b).

Considering the proof-of-principle studies presented, it seems likely that automated, AI-based assessments of therapeutic requirements will become a reliable component of management in retinal practice in the near future.

3.5. Prediction and prognosis

An exciting application scenario for AI methods is clearly to “foresee the future” based on pattern recognition in prior data. Precise prognostic tools would not only help to manage expectations of patients and doctors, improve the quality of care by providing optimal therapies but would also aid managing healthcare expenditure and introduce pragmatism into retinal therapy. The major targets for prediction include the functional outcomes after therapy and the future natural history course of a disease. However, in principle AI is able to produce predictive tools for any given target, provided that sufficient training data are available and that the task is per se solvable.

3.5.1. Prediction of visual acuity outcomes

The introduction of intravitreal anti-VEGF therapy is without doubt among the greatest achievements in retina in the last decades. However, ever-growing numbers of patients and interventions, substantial costs sometimes without a clearly visible benefit and difficulties in treatment planning constitute some of the challenges associated with the success story of anti-VEGFs. Furthermore, the development of therapeutic substances is limited because the available agents already show very high efficacy. Therefore, to differentiate one substance from another, an effective selection of study cohorts continues to rise in importance. AI definitively promises to solve several of these dilemmas by the means of validated, personalized prognostic tools. It may offer a precise, individualized forecast of visual outcomes after therapy, allow interpretation and ranking of imaging biomarkers, assist in the identification of new biomarkers and finally provide cohort stratification in substance development.

Neovascular AMD. The development of accurate systems to forecast the future development of visual acuity under intravitreal therapy represents one of the methodological break-throughs in AI research in recent years. The prediction of visual acuity outcomes is exceptionally relevant because patients with neovascular AMD in particular show substantial inter-individual variability in functional response to anti-VEGF therapy. A solid prognosis of vision outcomes after one or several years of therapy would likely lead to improved compliance by patients and better adherence by physicians to the appropriate treatment regimens. On the other hand, predictive tools may allow expensive, invasive therapy to be saved in individuals in whom any intervention would not be beneficial in the case of irreversible severe visual loss.

In the case of neovascular AMD, Schmidt-Erfurth et al. were the first to introduce a prognostic model that allowed forecasting of visual acuity outcomes after 12 months of anti-VEGF therapy in the setting of a randomized controlled trial within an error margin of 8.6 letters, i.e., close to the inter-session variability of a best-corrected visual acuity test (Schmidt-Erfurth et al., 2018a). The study was particularly comprehensive in including imaging-related biomarkers into the machine learning model (Figure 8). To allow a complete representation of OCT biomarkers in the modeling database, deep learning was used to extract a comprehensive set of known OCT biomarkers from the 3D OCT images acquired during the loading dose (Schlegl et al., 2018b; Zhang et al., 2014). These included, for instance, precise measurements of intraretinal fluid, subretinal fluid, pigment epithelial detachment and thicknesses of the individual retinal layers (Figure 8). The analysis resulted in over 200 spatially and temporally resolved variables to accurately represent each individual patient's retinal configuration. A random forest AI model was trained and validated using the known therapeutic response of over 600 patients receiving standardized ranibizumab therapy (in the context of a randomized controlled trial).

The model did not only predict individual visual acuity outcomes with an accuracy (R2) of 71% but its interpretation also allowed a comprehensive view into the specific biomarkers relevant for making the predictions. It confirmed that, among the current fluid-based markers, intraretinal cystoid fluid confers the most pronounced effect on visual acuity, i.e., a marked loss when large quantities of IRC are present in the fovea. The analysis highlighted, however, the surprisingly moderate overall correlation between retinal fluid on OCT and corresponding visual acuity, with a coefficient of determination of R2=0.21. Obviously, novel biomarkers must be sought for a better understanding of the mechanisms of vision loss in neovascular AMD. When analyzing the biomarkers for visual outcomes under therapy, the model clearly illustrated that the starting visual acuity of the individual patient and its initial response to therapy are almost exclusively the main determinants for final vision outcomes. Hence, research is being undertaken to examine additional biomarkers that could contribute to the prognosis of visual function (Schlegl et al., 2017). These preexisting, non-fluid-related markers may include a preexisting damage to neurosensory layers and RPE as seen by AI methods in intermediate AMD, and which may not recover easily additional fluid leakage (Schmidt-Erfurth et al., 2018b).

A similar predictive model was recently proposed based on electronic medical records contained in a mineable data warehouse (Rohm et al., 2018). The study did not consider complex OCT biomarkers but spatially resolved measurements of retinal thickness provided by the device segmentation software, which is prone to errors (Waldstein et al., 2015). Nevertheless, the investigators showed successful prediction of visual acuity outcomes after one year of real-world anti-VEGF therapy within an error margin of 8 letters using model developed and validated in 456 patients.

Diabetic macular edema and retinal vein occlusion. Analagous to the above-mentioned studies, other papers have offered prognostic AI models for DME and macular edema secondary to RVO. In the setting of diabetic macular disease, an AI model was developed based on data of the Protocol T study using patient-level information of 629 eyes and including advanced OCT image analysis (Gerendas et al., 2017). The study confirmed the significant importance of intraretinal cystoid fluid for visual acuity. However, the prediction of final vision outcomes was less precise at an R2 of 0.50 based on conventional biomarkers, highlighting again the need for novel biomarker searches. The prognostic value was highest for IRC resolution after the first injection.

In macular edema secondary to RVO, recent efforts have also used AI-based methods to analyze the predictive potential of OCT biomarkers. Work by Vogl et al. offered further insight by quantifying the visual damage conferred by retinal fluid, assigning 31 letters of BCVA loss for each mm^3 of intraretinal fluid in the foveal region (Vogl et al., 2017a). The model achieved a predictive accuracy at month 4 of R2 = 68% and an error margin of only 6 letters. A second paper used AI to segment the posterior vitreous boundary to diagnose the presence of a posterior vitreous detachment by means of unsupervised clustering (Waldstein et al., 2017). However, the study was not successful in identifying additional relevant biomarkers for vision outcomes, which highlights the limited conventional knowledge about disease-specific biomarkers and the need to further develop AI rankings of clinically relevant features.

3.5.2. Prediction of future natural disease course

Roughly a quarter of the population in industrialized countries over the age of 60 years is affected by early or intermediate dry AMD, representing one of the greatest pandemics in modern medicine. Early AMD is a chronically progressive disease characterized by a highly heterogeneous speed of advancement. It may remain at an early stage for the patient’s entire lifetime, without any relevant functional impairment, or may rapidly progress to advanced AMD, including CNV or GA with an associated massive functional morbidity. However, in clinical practice it can be very challenging to provide a robust prognosis with regards to progression speed, risk of advanced AMD and timing of the onset of advanced changes. At the moment, population-level studies provide risk scores. However, these may, obviously, not immediately translate to a given individual patient. This makes patient management difficult both because of the challenge in determining optimal follow-up intervals and because it leaves the patient worried with the uncertainty around his or her personal risk of future vision loss. Moreover, in terms of drug developments currently underway for dry AMD, it is imperative to have solid data on the risk and speed of the onset of advanced AMD. AI may allow the selection of study populations appropriately and may thus enable stratification of cohorts to include only patients in whom novel therapies would be likely to produce the measurable effect size given for the duration of the trial. In this context, AI models have been developed to provide a better understanding of the general manner of dry AMD progression and predictive models that deliver personalized risk prognosis for AMD conversion.

Drusen regression. To provide further insight into the main hallmark of early AMD, i.e., drusen and their development over time, researchers developed AI technology to model the growth and regression of drusen. Recent natural history data show that drusen exhibit a characteristic growth pattern with a cubic increase in volume over several years (Schlanitz et al., 2017). Once drusen volume reaches a critical threshold, sudden and rapid regression of drusen may occur. Drusen regression is closely associated with the onset of advanced AMD and development of CNV and/or GA in the exact area of the previously regressed drusen often occurs within a few months.

In the predictive model developed, it was possible to capture the usual growth pattern of drusen over time (Bogunovic et al., 2017a). 944 individual drusen (in 61 eyes) were identified by graph-cut analysis in the population studied; 26% of these drusen regressed within an observation period of up to 6 years. The AI model was successful in predicting the precise location and time of future drusen regression with an accuracy of up to 80% (Figure 10). Further research should be directed at the inclusion of healthy, elderly individuals to delve further into the differentiation between “healthy” and pathological ageing of the retina.

AMD conversion. The first AI model to provide a personalized prediction of AMD conversion was pioneered by a group at Stanford University in 2014 (de Sisternes et al., 2014). The model was trained and validated on quantitative features of drusen and retinal layers extracted from 330 eyes of 244 patients using automated segmentation algorithms. Random forest machine learning was used to create a statistical model that enabled determination of the individual disease progression risk within 5 years with an accuracy of 74% (Figure 11). However, differentiation between CNV and GA was not attempted and the work did not include pathognomonic features of AMD other than drusen and retinal layer thickness.

More recently, a new AI model to determine the risk of AMD conversion was proposed based on a larger dataset and a more comprehensive analysis of OCT biomarkers (Schmidt-Erfurth et al., 2018b). This study included data from 495 patients with CNV in one eye and intermediate AMD in the fellow eye, who were observed monthly during a randomized controlled study (providing ranibizumab therapy for the CNV eye). The analysis of this patient population offered particular value because of the high risk of progression in fellow eyes of patients with CNV. During the 24-month observation period, conversion to CNV was diagnosed in 114 eyes and development of GA in 45 eyes. Fully automated segmentation based on deep learning and graph cut was used to obtain a comprehensive representation of the retinal microanatomy, resulting in a volumetric quantification of drusen, hyperreflective foci, pseudodrusen and the individual retinal layers (Schlegl et al., 2018a). Based on these data, the investigators taught an AI model that was able to predict the development of CNV with an accuracy of 68% and the onset of GA with an accuracy of 80%, and for the first time enabled an a-priori differentiation between these two entities within advanced AMD. Most surprisingly, the characteristic key features leading to conversion towards SNV versus GA showed a distinctly different “signature” pattern supporting the notion that both are physiologically distinct pathways. Interestingly, genetic profiles were not relevant prognostic factors, and age only appeared as prognostic marker for GA, but nor for CNV.

An interpretation of the individual features considered by the random forest model offered revealing insights into the pathophysiology of AMD development (Figure 12). The two modes of conversion, i.e. CNV and GA, exhibited markedly different biomarkers that were considered relevant by the AI model. A high volume of drusen was the most important hallmark of disease progression in CNV. This is also supported by recent data showing subclinical macular neovascularization in intermediate AMD eyes, thus providing evidence for sub-RPE fluid exudation as an early sign of the onset of retinal exudation (de Oliveira Dias et al., 2018). By contrast, development of GA was mainly heralded by hyperreflective foci in the retina and loss of the outer neurosensory elements. Recent histopathology data supports the concept of hyperreflective foci representing migratory RPE cells that may be a sign of advancing RPE damage and disintegration (Curcio et al., 2017).

GA growth. Once a lesion of GA has developed, AI offers the opportunity to predict the direction and speed of future growth. Niu et al. proposed an AI model based on 29 patients and a mean observation period of 2.5 years (Niu et al., 2016). The model was able to foresee the future growth of GA lesions with a high accuracy, although comparison with a baseline (e.g. assuming linear, centrifugal GA growth at the established growth rates) was not provided (Figure 9). In this study, thinning of the outer retinal layers and the presence of reticular pseudodrusen were among the most important markers considered by the model. These early experimental results promise successful application of AI in analyzing GA, although refinement and validation in larger datasets should follow. Reliable AI models, particularly in the context of GA, will undoubtedly provide valuable support in counselling patients and in aiding the development of therapeutic interventions for GA.

4. Discussion

4.1. The potential of AI

The digital availability of information has already vastly transformed the practice of medicine. Modern physicians use Google and PubMed more frequently than text books to aid in diagnostic and therapeutic decisions (Kluwer, 2011). This is obvious as the breadth of medical knowledge and the speed of its development grow exponentially with the interval needed to double knowledge decreasing from 3.5 years in 2010 to a predicted 0.2 years by 2020 (Densen, 2011). Patients are getting older and are affected by more and more comorbidities. Diagnostic analyses offer enormous numbers of predictors which have to be integrated into prognostic equations. It does not surprise that most diagnostic tests in medicine come back negative and misdiagnosis is common (Care et al., 2015). Leveraging dramatic advances in computational power, digital voxel matrixes underlying retinal images become thousands of individual variables. Algorithms then cluster voxels into layers and contours, reconstruct 3D features from a 2D representation and ultimately learn pathognomonic patterns and disease categories. Such digital decision support is badly needed as even the frequent grading of DR is a complex task and agreement between clinicians certified for the task and manual, but standardized, reading center gradings in DR only reached consistency in 75% (Scott et al., 2008). Introduction of an automated algorithm for DR grading compared with retinal specialist gradings achieved substantial improvements in correct adjudication, including evident DR features such as microaneurysms (Krause et al., 2018). More sophisticated but relevant features such as photoreceptor disruption is not amenable to clinical evaluation but can be identified with an accuracy, sensitivity and specificity of more than 90% using automated detection on volumetric OCT (Wang et al., 2018). In metabolic disease including diabetes, multiethnicity may play an important role requiring huge datasets for validation and evaluation, e.g., 71,896 images/494,661 images only accessible by deep learning systems, reading an AUC of 0.94 for referable and vision-threating DR (Ting et al., 2017). AI using central telemedicine systems may also support poorly resourced services in areas where human expertise is missing.

Other medical fields have already highlighted the benefit of AI in their environment: CNN could detect tuberculosis in chest radiographs (Lakhani and Sundaram, 2017), melanoma from skin photographs more accurately than dermatologists (Esteva et al., 2017) and metastatic cells in lymph node samples more precisely than pathologists (Liu et al., 2017). Radiologists anticipate that the implementation of AI over the next decade will greatly improve the quality, value and depth of radiology's contribution to patient care and population health, and will revolutionize radiologists' workflow, as stated in the Canadian Association of Radiologists white paper on AI in Radiology (Tang et al., 2018). Retinology with its multimodal imaging modalities, high-resolution image quality, inexpensive and non-invasive approach should pioneer in the role of AI in medicine as diagnostic imaging is a major source of deep learning.

4.2. AI and personalized medicine

However, other specialties which appear less easily accessible to AI such as genetic counseling claim a goal to use AI to aid in identifying at-risk patients, generating differential diagnoses, improving efficiency in medical history collection and providing educational support for patients (Gordon et al., 2018) – a profile which may be copy-and-paste transferred to a disease such as AMD in the field of retina. Not to mention the approach of precision psychiatry in using machine learning for evidence-based psychiatry tailored to individual patients, objectively measurable endophenotypes allowing for early disease detection, individualized treatment selection and dosage adjustment to reduce the burden of disease – which are daily routine in medical retina (Bzdok and Meyer-Lindenberg, 2018). Personalized medicine is an urgent call in a healthcare system which cannot afford existing large redundancies together with a lack of recognition of individual conditions and needs. However, patient profiles are vastly different and difficult to recognize, even with a time-intensive physical examination, physician-doctor communication and expensive serologic or even genetic tests. Large initiatives have undertaken huge efforts to use genome-wide analysis of disease progression in AMD with the goal of assisting in early identification of high-risk individuals (Yan et al., 2018). Yet, it is questionable whether a genetic risk estimation will be as relevant as an individual imaging assessment using AI for a detailed individual and time-sensitive biomarker assessment which offers a signature profile in the conversion from early to advanced AMD (Schmidt-Erfurth et al., 2018b). Personalized medicine comes with the dilemma of time constraints in busy daily practices. Advances in electronic medical record analysis and comprehensive presentation of relevant previous and present features in particular have the potential to free the clinician to shift from disputing documentation and data-entry tasks derived from multiple sources, e.g., BCVA record, medication, fundus photography, angiography etc. to patient-focused activity. Proper interpretation and use of computerized data will depend as much on wise doctors as any other source of data has done in the past (Verghese et al., 2018). However, with the ability of AI to automate, e.g., in servicing electronic medical records, using speech recognition and image analysis, the physician will be able to extract the relevant features with a mouse click freeing-up more time for human-to-patient interactions, which will improve care and allow physicians to record and accurately register more individual phenotypes with added individual nuance (Halpern et al., 2016). Eric Topol, the pioneer in digital medicine, refers to the digital revolution in medicine as “the creative destruction of medicine” (Topol, 2011). He also highlights the socioeconomic opportunity of AI-guided medicine in his book “The patient will see you now: The future of medicine is in your hands”. With the advent of patient-accessible automated scanners, individuals can take advantage of screening procedures without the need to wait for a doctor’s appointment. Physicians can also spread their knowledge across disciplines, as AI-based systems brings diagnostic expertise in retina into primary care in an interdisciplinary way. The high resolution of retinal imaging in particular enables the physician to assess human health at an unprecedented level. The Google-Project extracted highly personalized data such as sex, age, blood pressure, HbA1c and smoking history from a single digital color photograph of the retina, far beyond ophthalmological relevance (Poplin et al., 2018). This capability brings ophthalmology/retinology into the focus of high-definition medicine as a dynamic assessment, management and understanding of an individual’s health over life-time. Strategies of high-definition medicine include defining a personal therapy and establishing a continuously improving learning healthcare system (Torkamani et al., 2017).

4.3. Challenges in AI-based retina

Data access and therefore big data sharing are quintessential issues in machine/deep learning and neural networks are intrinsically “data hungry”. The public availability of ImageNet in 2009 catalyzed AI and is still the base of retina-based image analyses (Deng et al., 2009). Open access to scientific data has become a prominent topic on the global research agenda. While the rise of open access policies is fundamentally changing the academic landscape, it is reigniting the conversation around adequate policies to protect scientific intellectual property. In 2015, Hahnel referred to “the open academic tidal wave” describing the transition from open access to scholarly papers of publicly funded research, to access of all digital outputs, to mandated and enforced access to all digital outputs of publicly funded research (Hahnel, 2015) (Figure 19).

Open access to research data, which largely includes images in retinal research, is made mandatory by important funding agencies such as the National Eye Institute (NEI) and the Wellcome Trust, and has the potential to bring retinal research to prolific horizons. The UK Biobank initiative is an excellent example of open access retinal imaging. This biobank aggregated self-reported disease questionnaires and physical and eye examinations, including macular SD-OCT scans, from about 67,000 individuals aged 40 to 69 years for systematic analysis of macular thickness and associations with RPE measurements (Keane et al., 2016; Ko et al., 2017; Patel et al., 2016).

Independent of the value of open data access for research, medical data is fundamentally and legally different. The NHS's initiative to share identifiable clinical data of 1.6 million patients with Google/Deep mind (with the goal to develop an app to monitor patients at risk of acute kidney injury) has raised substantial questions about data confidentiality, particularly as this process was not made public until investigative journalism actively interfered (The Guardian, 2017). The fear is obviously that algorithms based on confidential NHS records will seed an entirely new industry in AI-based technology. Questions regarding privacy protection are particularly sensitive in retinal imaging as anonymization is not completely achievable due to the individual nature of the retinal vasculature which provides a fingerprint-like individual feature. The fact is, it is not possible to completely anonymize any medical images, whether they are MRIs of the brain or ophthalmic images. For this reason, data protection experts and ethical bodies now refer to "de-identification" or "de-personalization" of medical images. They also require that, given the challenges in complete anonymization of any images, that appropriate safe guards be put in place to further reduce the theoretical risks of re-identification. In addition, questions such as data ownership, rights to intellectual property and big profits created from public funding become more and more virulent (Beam and Kohane, 2018). Requirements of data protection and pseudonymization for safe data transmission and redundant privacy-compliant storage with disaster recovery plans are hugely expensive as imaging datasets are big data on a per patient level. Cyberattacks may jeopardize automated screening tools with so-called adversarial samples against deep learning systems which are otherwise invisible to the human expert. The healthcare economy and its multiple incentives make it particularly sensitive to fraud. The challenge of incorporating ethics into data technologies is formidable. This is in part because it requires overcoming a century-long ethos of data science: develop first, question later; datafication first, regulation afterwards (Koopman, 2018). This criticism implies that innovative research often proceeds proactively with presenting paradigm-shifting discoveries, while a comprehensive evaluation of all possible side-effects and limitations follows subsequently when the community has the opportunity to embrace the change on a larger scale and a real-world setting. Particularly novel means of big data analyses have to cope with this phenomenon as highlighted in an exemplary manner by the Google/NHS project which has initiated intensive legal ramifications subsequently.

The other "elephant in the room" is the black-box phenomenon. In deep learning, it is challenging to understand how exactly a neural network reaches a particular decision, or to identify which exact features it utilizes (see Section 2.7). As AI already outperformed human expertise, how can the results of AI-based algorithms be properly understood by clinicians and researchers? How can we ensure the reliability of algorithms, if we cannot understand how they operate? Potential solutions to this problem are multi-step algorithms that first detect certain clinically known features (using deep learning) and then predict or classify based on these features. However, the value of an end-to-end approach with the potential for a higher accuracy and the discovery of new markers is obviously lost in this trade-off.

Another limitation represents the possibility of inherent bias in AI that has to be recognized. The representative value has to be evaluated. In many cases, the analysis of big data goes beyond direct human intelligence (Balthazar et al., 2018). Algorithms learn from data compiled in current clinical practice. Therefore, AI-based algorithms in anti-VEGF trails strongly rely on the nature study protocols and behavioral procedures including mandatory retreatments whenever intra- or subretinal fluid becomes apparent, potentially leading to overtreatment (Schmidt-Erfurth et al., 2018a). Despite the golden rule in anti-VEGF therapy to rigorously eliminate fluid from the neurosensory retina, the correlation between retinal fluid and retinal function was found to extremely low at an R^2 of 0.23. AI-based outcomes necessarily require comprehensive intellectual validation based on clinical expertise but may also open the horizon for novel insight into the pathophysiology of retinal disease. In real-world data analyses, algorithms that learn from human decisions are particularly likely to reiterate human errors (Obermeyer and Lee, 2017). Although machine-learning methods are especially suited to making predictions based on existing data, precise predictions about the distant future are often fundamentally inaccurate. The rise and fall of Google Flu is a reminder of the complexity of forecasting as is the insufficiency of treat-and-extend data. With fast changing diagnostic and therapeutic paradigms, previous data sets have a short survival and the relevance of clinical data decays with an effective “half-life” of about 4 months (Chen et al., 2017). Although predictive algorithms are unable to provide absolute medical certainty, they may strongly improve allocations of stressed healthcare resources by helping to plan large-scale patient care, comparing the efficiency of therapeutic substances and suggesting sound treatment indications, which is a future must in the pharmacological era of retinal therapy (Chen and Asch, 2017). Retinologists are called to reorganize their specialty according to their patients’ needs. They need to defend their field against destructive reimbursement policies which lead to miserable outcomes in the real world (LUMINOUS) and to navigate soundly between the goals of improving health and generating profit. Modern intelligent tools can support this task. If understood in depth and applied with expertise, AI offers the unique opportunity to establish a collective medical mind combining published research, big data analysis and individual expertise with the tenets of professional ethics.

The black-box phenomena is particularly intrinsic to daily routine as digital imaging focuses on subclinical biomarkers such as hyperreflective foci, deep capillary plexus and other features beyond clinically visible correlates. Hence, AI-based detection and integration is not an alternative but a necessity. A collaborative approach is the only path towards meaningful insight by big data analysis which may strengthen the entire field substantially and raise overall quality.

Finally, it is important to point out that most AI-based applications in medicine are still in the translational stage and have not yet demonstrated their benefit in clinical trials. However, the authors believe that it is merely a matter of time until this hurdle will be successfully taken.

From a visionary perspective, AI in retina may appear rather “organic” as human visual perception works in a similar way to feature recognition by AI: an image is projected to the photoreceptors of the retina, representing the first neuronal layer, which feeds it forward to neurons in subsequent neurosensory layers, which then forward the visual signals to multiple connected neuronal networks in the visual cortex and associated areas in the brain that process visual stimuli simultaneously and in real time. Human visual perception is also established by learning and combining images using labels, rather like self-teaching systems in machine learning. The challenge is now to integrate such a highly developed system into our profession.

References

Abdelfattah, N.S., Zhang, H., Boyer, D.S., Rosenfeld, P.J., Feuer, W.J., Gregori, G., Sadda, S.R., 2016. Drusen Volume as a Predictor of Disease Progression in Patients With Late Age-Related Macular Degeneration in the Fellow Eye. Investigative Ophthalmology & Visual Science 57, 1839-1846.

Abramoff, M.D., Folk, J.C., Han, D.P., Walker, J.D., Williams, D.F., Russell, S.R., Massin, P., Cochener, B., Gain, P., Tang, L., Lamard, M., Moga, D.C., Quellec, G., Niemeijer, M., 2013. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol 131, 351-357.

Abramoff, M.D., Garvin, M.K., Sonka, M., 2010. Retinal imaging and image analysis. IEEE Rev Biomed Eng 3, 169-208.

Abramoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M., 2016. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. Invest Ophthalmol Vis Sci 57, 5200-5206.

Adhi, M., Duker, J.S., 2013. Optical coherence tomography--current and future applications. Curr Opin Ophthalmol 24, 213-221.

Al-Bander, B., Al-Nuaimy, W., Williams, B.M., Zheng, Y., 2018. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. Biomedical Signal Processing and Control 40, 91-101.

Allam, A.M.N., Youssif, A.A.-H., Ghalwash, A.Z., 2017. Segmentation of Exudates via Color-based K-means Clustering and Statistical-based Thresholding. Journal of Computer Science 13, 524-536.

Alsaih, K., Lemaitre, G., Rastgoo, M., Massich, J., Sidibé, D., Meriaudeau, F., 2017. Machine learning techniques for diabetic macular edema (DME) classification on SD-OCT images. BioMedical Engineering OnLine 16, 68.

Arnold, J.J., Markey, C.M., Kurstjens, N.P., Guymer, R.H., 2016. The role of sub-retinal fluid in determining treatment outcomes in patients with neovascular age-related macular degeneration--a phase IV randomised clinical trial with ranibizumab: the FLUID study. BMC Ophthalmol 16, 31.

Ataer-Cansizoglu, E., Bolon-Canedo, V., Campbell, J.P., Bozkurt, A., Erdogmus, D., Kalpathy-Cramer, J., Patel, S., Jonas, K., Chan, R.V.P., Ostmo, S., Chiang, M.F., on behalf of the i, R.O.P.R.C., 2015. Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the “i-ROP” System and Image Features Associated With Expert Diagnosis. Translational Vision Science & Technology 4, 5.

Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H., 2017. Improving Palliative Care with Deep Learning, in: IEEE (Ed.), IEEE International Conference on Bioinformatics and Biomedicine 2017.

Balthazar, P., Harri, P., Prater, A., Safdar, N.M., 2018. Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. J Am Coll Radiol 15, 580-586.

Baxt, W.G., 1991. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 115, 843-848.

Beam, A.L., Kohane, I.S., 2018. Big Data and Machine Learning in Health Care. JAMA 319, 1317-1318.

Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T., Boss, A., 2017. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Invest Radiol 52, 434-440.

Bogunovic, H., Kwon, Y.H., Rashid, A., Lee, K., Critser, D.B., Garvin, M.K., Sonka, M., Abramoff, M.D., 2014. Relationships of retinal structure and humphrey 24-2 visual field thresholds in patients with glaucoma. Invest Ophthalmol Vis Sci 56, 259-271.

Bogunovic, H., Montuoro, A., Baratsits, M., Karantonis, M.G., Waldstein, S.M., Schlanitz, F., Schmidt-Erfurth, U., 2017a. Machine Learning of the Progression of Intermediate Age-Related Macular Degeneration Based on OCT Imaging. Invest Ophthalmol Vis Sci 58, BIO141-BIO150.

Bogunovic, H., Waldstein, S.M., Sadeghipour, A., Gerendas, B.S., Schmidt-Erfurth, U., 2018. Artificial intelligence to predict optimal retreatment intervals in treat-and-extend anti-VEGF therapy Invest Ophthalmol Vis Sci, ARVO E-Abstract.

Bogunovic, H., Waldstein, S.M., Schlegl, T., Langs, G., Sadeghipour, A., Liu, X., Gerendas, B.S., Osborne, A., Schmidt-Erfurth, U., 2017b. Prediction of Anti-VEGF Treatment Requirements in Neovascular AMD Using a Machine Learning Approach. Invest Ophthalmol Vis Sci 58, 3240-3248.

Bowd, C., Hao, J., Tavares, I.M., Medeiros, F.A., Zangwill, L.M., Lee, T.-W., Sample, P.A., Weinreb, R.N., Goldbaum, M.H., 2008. Bayesian Machine Learning Classifiers for Combining Structural and Functional Measurements to Classify Healthy and Glaucomatous Eyes. Investigative Ophthalmology & Visual Science 49, 945-953.

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 37, 819-835.

Breger, A., Ehler, M., Bogunovic, H., Waldstein, S.M., Philip, A.M., Schmidt-Erfurth, U., Gerendas, B.S., 2017. Supervised learning and dimension reduction techniques for quantification of retinal fluid in optical coherence tomography images. Eye (Lond) 31, 1212-1220.

Browning, D.J., Glassman, A.R., Aiello, L.P., Beck, R.W., Brown, D.M., Fong, D.S., Bressler, N.M., Danis, R.P., Kinyoun, J.L., Nguyen, Q.D., Bhavsar, A.R., Gottlieb, J., Pieramici, D.J., Rauser, M.E., Apte, R.S., Lim, J.I., Miskala, P.H., 2007. Relationship between optical coherence tomography-measured central retinal thickness and visual acuity in diabetic macular edema. Ophthalmology 114, 525-536.

Buchanan, B.G., Shortliffe, E.H., 1984. Rule-based expert systems: The mycin experiments of the stanford heuristic programming project. USC/Information Sciences Institute, Marina del Rey, CA 90292, U.S.A.

Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., Bressler, N.M., 2017. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. JAMA Ophthalmol 135, 1170-1176.

Busbee, B.G., Ho, A.C., Brown, D.M., Heier, J.S., Suner, I.J., Li, Z., Rubio, R.G., Lai, P., 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.

Bzdok, D., Meyer-Lindenberg, A., 2018. Machine Learning for Precision Psychiatry: Opportunities and Challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 3, 223-230.

Campbell, J.P., Ataer-Cansizoglu, E., Bolon-Canedo, V., Bozkurt, A., Erdogmus, D., Kalpathy-Cramer, J., Patel, S.N., Reynolds, J.D., Horowitz, J., Hutcheson, K., Shapiro, M., Repka, M.X., Ferrone, P., Drenser, K., Martinez-Castellanos, M.A., Ostmo, S., Jonas, K., Chan, R.V.P., Chiang, M.F., on behalf of the i, R.O.P.r.c., 2016. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. JAMA ophthalmology 134, 651-657.

Care, C.o.D.E.i.H., Services, B.o.H.C., Medicine, I.o., The National Academies of Sciences, E., and Medicine, 2015. Improving Diagnosis in Health Care, in: Balogh, E.P., Miller, B.T., Ball, J.R. (Eds.), Improving Diagnosis in Health Care. National Academies Press, Washington (DC).

Chakravarthy, U., Goldenberg, D., Young, G., Havilio, M., Rafaeli, O., Benyamini, G., Loewenstein, A., 2016. Automated Identification of Lesion Activity in Neovascular Age-Related Macular Degeneration. Ophthalmology 123, 1731-1736.

Chen, J.H., Alagappan, M., Goldstein, M.K., Asch, S.M., Altman, R.B., 2017. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform 102, 71-79.

Chen, J.H., Asch, S.M., 2017. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 376, 2507-2509.

Chen, Q., Leng, T., Zheng, L., Kutzscher, L., Ma, J., de Sisternes, L., Rubin, D.L., 2013. Automated Drusen Segmentation and Quantification in SD-OCT Images. Medical image analysis 17, 1058-1072.

Chiu, C.J., Mitchell, P., Klein, R., Klein, B.E., Chang, M.L., Gensler, G., Taylor, A., 2014. A risk score for the prediction of advanced age-related macular degeneration: development and validation in 2 prospective cohorts. Ophthalmology 121, 1421-1427.

Choi, J.Y., Yoo, T.K., Seo, J.G., Kwak, J., Um, T.T., Rim, T.H., 2017. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PloS one 12, e0187336.

Curcio, C.A., Zanzottera, E.C., Ach, T., Balaratnasingam, C., Freund, K.B., 2017. Activated Retinal Pigment Epithelium, an Optical Coherence Tomography Biomarker for Progression in Age-Related Macular Degeneration. Invest Ophthalmol Vis Sci 58, BIO211-BIO226.

de Oliveira Dias, J.R., Zhang, Q., Garcia, J.M.B., Zheng, F., Motulsky, E.H., Roisman, L., Miller, A., Chen, C.L., Kubach, S., de Sisternes, L., Durbin, M.K., Feuer, W., Wang, R.K., Gregori, G., Rosenfeld, P.J., 2018. Natural History of Subclinical Neovascularization in Nonexudative Age-Related Macular Degeneration Using Swept-Source OCT Angiography. Ophthalmology 125, 255-266.

de Sisternes, L., Jonna, G., Greven, M.A., Chen, Q., Leng, T., Rubin, D.L., 2017. Individual Drusen Segmentation and Repeatability and Reproducibility of Their Automated Quantification in Optical Coherence Tomography Images. Translational Vision Science & Technology 6, 12.

de Sisternes, L., Simon, N., Tibshirani, R., Leng, T., Rubin, D.L., 2014. Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression. Invest Ophthalmol Vis Sci 55, 7093-7103.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 248-255.

Densen, P., 2011. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 122, 48-58.

Devalla, S.K., Chin, K.S., Mari, J.-M., Tun, T.A., Strouthidis, N.G., Aung, T., Thiéry, A.H., Girard, M.J.A., 2018. A Deep Learning Approach to Digitally Stain Optical Coherence Tomography Images of the Optic Nerve Head. Investigative Ophthalmology & Visual Science 59, 63-74.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., the, C.C., Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.J., Heng, P.A., Hass, C., Bruni, E., Wong, Q., Halici, U., Oner, M.U., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., Venancio, R., 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199-2210.

ElTanboly, A., Ismail, M., Shalaby, A., Switala, A., El-Baz, A., Schaal, S., Gimel'farb, G., El-Azab, M., 2017. A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images. Med Phys 44, 914-923.

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, 115-118.

Fan, Z., Rong, Y., Cai, X., Lu, J., Li, W., Lin, H., Chen, X., 2018. Optic Disk Detection in Fundus Image Based on Structured Learning. IEEE Journal of Biomedical and Health Informatics 22, 224-234.

Farsiu, S., Chiu, S.J., O'Connell, R.V., Folgar, F.A., Yuan, E., Izatt, J.A., Toth, C.A., Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study, G., 2014. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology 121, 162-172.

Fatima, K.N., Hassan, T., Akram, M.U., Akhtar, M., Butt, W.H., 2017. Fully automated diagnosis of papilledema through robust extraction of vascular patterns and ocular pathology from fundus photographs. Biomed Opt Express 8, 1005-1024.
Feeny, A.K., Tadarati, M., Freund, D.E., Bressler, N.M., Burlina, P., 2015. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. Comput Biol Med 65, 124-136.
Fei, X., Zhao, J., Zhao, H., Yun, D., Zhang, Y., 2017. Deblurring adaptive optics retinal images using deep convolutional neural networks. Biomed Opt Express 8, 5675-5687.
Figueiredo, I.N., Kumar, S., Oliveira, C.M., Ramos, J.D., Engquist, B., 2015. Automated lesion detectors in retinal fundus images. Computers in Biology and Medicine 66, 47-65.
Fragiotta, S., Rossi, T., Cutini, A., Grenga, P.L., Vingolo, E.M., 2018. PREDICTIVE FACTORS FOR DEVELOPMENT OF NEOVASCULAR AGE-RELATED MACULAR DEGENERATION: A Spectral-Domain Optical Coherence Tomography Study. Retina 38, 245-252.
Freund, K.B., Korobelnik, J.-F., Devenyi, R., Framme, C., Galic, J., Herbert, E., Hoerauf, H., Lanzetta, P., Michels, S., Mitchell, P., Monés, J., Regillo, C., Tadayoni, R., Talks, J., Wolf, S., 2015. TREAT-AND-EXTEND REGIMENS WITH ANTI-VEGF AGENTS IN RETINAL DISEASES: A Literature Review and Consensus Recommendations. RETINA 35, 1489-1506.
García, M., Sánchez, C.I., López, M.I., Abásolo, D., Hornero, R., 2009a. Neural network based detection of hard exudates in retinal images. Computer Methods and Programs in Biomedicine 93, 9-19.
García, M., Sánchez, C.I., Poza, J., López, M.I., Hornero, R., 2009b. Detection of Hard Exudates in Retinal Images Using a Radial Basis Function Classifier. Annals of Biomedical Engineering 37, 1448-1463.
Gardner, G.G., Keating, D., Williamson, T.H., Elliott, A.T., 1996. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. The British journal of ophthalmology 80, 940-944.
Gargeya, R., Leng, T., 2017. Automated Identification of Diabetic Retinopathy Using Deep Learning. Ophthalmology 124, 962-969.
GeethaRamani, R., Balasubramanian, L., 2018. Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening. Computer Methods and Programs in Biomedicine 160, 153-163.
Gerendas, B.S., Bogunovic, H., Sadeghipour, A., Schlegl, T., Langs, G., Waldstein, S.M., Schmidt-Erfurth, U., 2017. Computational image analysis for prognosis determination in DME. Vision Res.
Gerendas, B.S., Prager, S., Deak, G., Simader, C., Lammer, J., Waldstein, S.M., Guerin, T., Kundi, M., Schmidt-Erfurth, U.M., 2018. Predictive imaging biomarkers relevant for functional and anatomical outcomes during ranibizumab therapy of diabetic macular oedema. The British journal of ophthalmology 102, 195-203.
Gordon, E.S., Babu, D., Laney, D.A., 2018. The future is now: Technology's impact on the practice of genetic counseling. Am J Med Genet C Semin Med Genet 178, 15-23.
Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R., 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama.
Guo, Z., Kwon, Y.H., Lee, K., Wang, K., Wahle, A., Alward, W.L.M., Fingert, J.H., Bettis, D.I., Johnson, C.A., Garvin, M.K., Sonka, M., Abramoff, M.D., 2017. Optical Coherence Tomography Analysis Based Prediction of Humphrey 24-2 Visual Field Thresholds in Patients With Glaucoma. Invest Ophthalmol Vis Sci 58, 3975-3985.
Hahnel, M., 2015. 2015 - The year of open data mandates.
Haleem, M.S., Han, L., Hemert, J.v., Fleming, A., Pasquale, L.R., Silva, P.S., Song, B.J., Aiello, L.P., 2016. Regional Image Features Model for Automatic Classification between Normal and Glaucoma in Fundus and Scanning Laser Ophthalmoscopy (SLO) Images. Journal of Medical Systems 40, 132.

Haleem, M.S., Han, L., Hemert, J.v., Li, B., Fleming, A., Pasquale, L.R., Song, B.J., 2017. A Novel Adaptive Deformable Model for Automated Optic Disc and Cup Segmentation to Aid Glaucoma Diagnosis. Journal of Medical Systems 42, 20.

Halpern, Y., Horng, S., Choi, Y., Sontag, D., 2016. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc 23, 731-740.

Han, S.S., Park, G.H., Lim, W., Kim, M.S., Na, J.I., Park, I., Chang, S.E., 2018. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PloS one 13, e0191493.

Harangi, B., Hajdu, A., 2014. Automatic exudate detection by fusing multiple active contours and regionwise classification. Computers in Biology and Medicine 54, 156-171.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

He, Y., Carass, A., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2018. Topology guaranteed segmentation of the human retina from OCT using convolutional neural networks. arXiv:1803.05120 [cs].

Heimes, B., Schick, T., Brinkmann, C.K., Wiedon, A., Haegele, B., Kirchhof, B., Holz, F.G., Pauleikhoff, D., Ziemssen, F., Liakopoulos, S., Spital, G., Schmitz-Valckenberg, S., 2016. Design des ORCA-Moduls der OCEAN-Studie. Der Ophthalmologe 113, 570-580.

Issac, A., Partha Sarathi, M., Dutta, M.K., 2015. An adaptive threshold based image processing technique for improved glaucoma detection and classification. Computer Methods and Programs in Biomedicine 122, 229-244.

Ji, Z., Chen, Q., Niu, S., Leng, T., Rubin, D.L., 2018. Beyond Retinal Layers: A Deep Voting Model for Automated Geographic Atrophy Segmentation in SD-OCT Images. Transl Vis Sci Technol 7, 1.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2, 230-243.

Kao, E.-F., Lin, P.-C., Chou, M.-C., Jaw, T.-S., Liu, G.-C., 2014. Automated detection of fovea in fundus images based on vessel-free zone and adaptive Gaussian template. Computer Methods and Programs in Biomedicine 117, 92-103.

Karri, S.P., Chakraborty, D., Chatterjee, J., 2017. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomed Opt Express 8, 579-592.

Kaur, J., Mittal, D., 2018. A generalized method for the segmentation of exudates from pathological retinal fundus images. Biocybernetics and Biomedical Engineering 38, 27-53.

Keane, P.A., Grossi, C.M., Foster, P.J., Yang, Q., Reisman, C.A., Chan, K., Peto, T., Thomas, D., Patel, P.J., Consortium, U.K.B.E.V., 2016. Optical Coherence Tomography in the UK Biobank Study - Rapid Automated Analysis of Retinal Thickness for Large Population-Based Studies. PloS one 11, e0164095.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M.Y.L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A.N., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K., 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172, 1122-1131 e1129.

Khalid, S., Akram, M.U., Hassan, T., Jameel, A., Khalil, T., 2017. Automated Segmentation and Quantification of Drusen in Fundus and Optical Coherence Tomography Images for Detection of ARMD. Journal of Digital Imaging.

Kharghanian, R., Ahmadyfard, A., 2012. Retinal Blood Vessel Segmentation Using Gabor Wavelet and Line Operator. International Journal of Machine Learning and Computing, 593-597.

Kim, S.J., Cho, K.J., Oh, S., 2017. Development of machine learning models for diagnosis of glaucoma. PloS one 12, e0177726.

Kluwer, W., 2011. Wolters Kluwer Health 2011 point-of-care-survey: Physicians face disconnects at point-of-care.

Ko, F., Foster, P.J., Strouthidis, N.G., Shweikh, Y., Yang, Q., Reisman, C.A., Muthy, Z.A., Chakravarthy, U., Lotery, A.J., Keane, P.A., Tufail, A., Grossi, C.M., Patel, P.J., Eye, U.K.B., Vision, C., 2017. Associations with Retinal Pigment Epithelium Thickness Measures in a Large Cohort: Results from the UK Biobank. Ophthalmology 124, 105-117.

Koopman, C., 2018. How Democracy Can Survive Big Data, The New York Times.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R., 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology.

Lakhani, P., Sundaram, B., 2017. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 284, 574-582.

Lang, A., Carass, A., Hauser, M., Sotirchos, E.S., Calabresi, P.A., Ying, H.S., Prince, J.L., 2013. Retinal layer segmentation of macular OCT images using boundary classification. Biomedical Optics Express 4, 1133-1152.

Larson, D.B., Chen, M.C., Lungren, M.P., Halabi, S.S., Stence, N.V., Langlotz, C.P., 2018. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology 287, 313-322.

Lawrence, M.G., 2004. The accuracy of digital-video retinal imaging to screen for diabetic retinopathy: an analysis of two digital-video retinal imaging systems using standard stereoscopic seven-field photography and dilated clinical examination as reference standards. Transactions of the American Ophthalmological Society 102, 321-340.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444.

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

Lee, C.S., Baughman, D.M., Lee, A.Y., 2017a. Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. Ophthalmology Retina 1, 322-327.

Lee, C.S., Tyring, A.J., Deruyter, N.P., Wu, Y., Rokem, A., Lee, A.Y., 2017b. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomedical Optics Express 8, 3440-3448.

Lee, C.S., Tyring, A.J., Wu, Y., Xiao, S., Rokem, A.S., Deruyter, N.P., Zhang, Q., Tufail, A., Wang, R.K., Lee, A.Y., 2018a. Generating perfusion maps from structural optical coherence tomography with artificial intelligence. bioRxiv: 271346.

Lee, H., Kang, K.E., Chung, H., Kim, H.C., 2018b. Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration. American Journal of Ophthalmology.

Lee, R., Wong, T.Y., Sabanayagam, C., 2015. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis (Lond) 2, 17.

Liefers, B., Venhuizen, F.G., Schreur, V., van Ginneken, B., Hoyng, C., Fauser, S., Theelen, T., Sánchez, C.I., 2017. Automatic detection of the foveal center in optical coherence tomography. Biomedical Optics Express 8, 5160-5178.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sanchez, C.I., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60-88.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C., 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv.org, arXiv:1703.02442.

Liu, Y.-Y., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J.S., Rehg, J.M., 2011a. Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. Medical Image Analysis 15, 748-759.

Liu, Y.-Y., Ishikawa, H., Chen, M., Wollstein, G., Duker, J.S., Fujimoto, J.G., Schuman, J.S., Rehg, J.M., 2011b. Computerized Macular Pathology Diagnosis in Spectral Domain Optical Coherence Tomography Scans Based on Multi-Scale Texture and Shape Features. Investigative Ophthalmology & Visual Science.

Maclin, P.S., Dempsey, J., Brooks, J., Rand, J., 1991. Using neural networks to diagnose cancer. J Med Syst 15, 11-19.

Mehta, H., Tufail, A., Daien, V., Lee, A.Y., Nguyen, V., Ozturk, M., Barthelmes, D., Gillies, M.C., 2018. Real-world outcomes in patients with neovascular age-related macular degeneration treated with intravitreal vascular endothelial growth factor inhibitors. Prog Retin Eye Res.

Memari, N., Ramli, A.R., Bin Saripan, M.I., Mashohor, S., Moghbel, M., 2017. Supervised retinal vessel segmentation from color fundus images based on matched filtering and AdaBoost classifier. PloS one 12.

Miri, M.S., Abràmoff, M.D., Kwon, Y.H., Sonka, M., Garvin, M.K., 2017. A machine-learning graph-based approach for 3D segmentation of Bruch's membrane opening from glaucomatous SD-OCT volumes. Medical Image Analysis 39, 206-217.

Miri, M.S., Abràmoff, M.D., Lee, K., Niemeijer, M., Wang, J.K., Kwon, Y.H., Garvin, M.K., 2015. Multimodal Segmentation of Optic Disc and Cup From SD-OCT and Color Fundus Photographs Using a Machine-Learning Graph-Based Approach. IEEE Transactions on Medical Imaging 34, 1854-1866.

Moccia, S., De Momi, E., El Hadji, S., Mattos, L.S., 2018. Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. Computer Methods and Programs in Biomedicine 158, 71-91.

Molina-Casado, J.M., Carmona, E.J., García-Feijoó, J., 2017. Fast detection of the main anatomical structures in digital retinal images based on intra- and inter-structure relational knowledge. Computer Methods and Programs in Biomedicine 149, 55-68.

Montuoro, A., Waldstein, S.M., Gerendas, B.S., Schmidt-Erfurth, U., Bogunovic, H., 2017. Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context. Biomed Opt Express 8, 1874-1888.

Murakami, Y., Jain, A., Silva, R.A., Lad, E.M., Gandhi, J., Moshfeghi, D.M., 2008. Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP): 12-month experience with telemedicine screening. The British journal of ophthalmology 92, 1456-1460.

Niu, S., de Sisternes, L., Chen, Q., Rubin, D.L., Leng, T., 2016. Fully Automated Prediction of Geographic Atrophy Growth Using Quantitative Spectral-Domain Optical Coherence Tomography Biomarkers. Ophthalmology 123, 1737-1750.

Obermeyer, Z., Emanuel, E.J., 2016. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 375, 1216-1219.

Obermeyer, Z., Lee, T.H., 2017. Lost in Thought - The Limits of the Human Mind and the Future of Medicine. N Engl J Med 377, 1209-1211.

Ohsugi, H., Tabuchi, H., Enno, H., Ishitobi, N., 2017. Accuracy of deep learning, a machine-learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment. Sci Rep 7, 9425.

Patel, P.J., Foster, P.J., Grossi, C.M., Keane, P.A., Ko, F., Lotery, A., Peto, T., Reisman, C.A., Strouthidis, N.G., Yang, Q., Eyes, U.K.B., Vision, C., 2016. Spectral-Domain Optical Coherence Tomography Imaging in 67 321 Adults: Associations with Macular Thickness in the UK Biobank Study. Ophthalmology 123, 829-840.

Penha, F.M., Gregori, G., Garcia Filho, C.A., Yehoshua, Z., Feuer, W.J., Rosenfeld, P.J., 2013. Quantitative changes in retinal pigment epithelial detachments as a predictor for retreatment with anti-VEGF therapy. Retina 33, 459-466.

Pennington, K.L., DeAngelis, M.M., 2016. Epidemiology of age-related macular degeneration (AMD): associations with cardiovascular disease phenotypes and lipid factors. Eye Vis (Lond) 3, 34.

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, 158-164.

Prahs, P., Radeck, V., Mayer, C., Cvetkov, Y., Cvetkova, N., Helbig, H., Märker, D., 2018. OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. Graefe's Archive for Clinical and Experimental Ophthalmology 256, 91-98.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J., 2018. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 1, 18.

Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y., 2017a. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. arXiv.org, arXiv:1707.01836.

Rajpurkar, P. , Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., Ng, A. Y., 2017b CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv.org, arXiv:1711.05225

Relan, D., MacGillivray, T., Ballerini, L., Trucco, E., 2014. Automatic retinal vessel classification using a Least Square-Support Vector Machine in VAMPIRE. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2014, 142-145.

Ren, X., Zheng, Y., Zhao, Y., Luo, C., Wang, H., Lian, J., He, Y., 2018. Drusen Segmentation From Retinal Images via Supervised Feature Learning. IEEE Access 6, 2952-2961.

Rohm, M., Tresp, V., Muller, 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.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, Vol.9351: 234-241 .

Roy, A.G., Conjeti, S., Karri, S.P.K., Sheet, D., Katouzian, A., Wachinger, C., Navab, N., 2017. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomedical Optics Express 8, 3627-3642.

Rubin, D.L., Chen, Q., Leng, T., Zheng, L.L., Kutzscher, L., de Sisternes, L., 2013. Improving drusen visualization on projection images in optical coherence tomography. Ophthalmology 120, 644-644.e642.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211-252.

Russell, S.J., Norvig, P., 1995. Artificial Intelligence: A Modern Approach. Prentice Hall.

Sayegh, R.G., Sacu, S., Dunavolgyi, R., Kroh, M.E., Roberts, P., Mitsch, C., Montuoro, A., Ehrenmuller, M., Schmidt-Erfurth, U., 2017. Geographic Atrophy and Foveal-Sparing Changes Related to Visual Acuity in Patients With Dry Age-Related Macular Degeneration Over Time. Am J Ophthalmol 179, 118-128.

Schlanitz, F.G., Baumann, B., Kundi, M., Sacu, S., Baratsits, M., Scheschy, U., Shahlaee, A., Mittermuller, T.J., Montuoro, A., Roberts, P., Pircher, M., Hitzenberger, C.K., Schmidt-Erfurth, U., 2017. Drusen volume development over time and its relevance to the course of age-related macular degeneration. The British journal of ophthalmology 101, 198-203.

Schlegl, T., Bogunovic, H., Klimscha, S., Seeboeck, P., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2018a. Fully Automated Segmentation of Hyperreflective Foci in Optical Coherence Tomography Images. arXiv.org, arXiv:1805.03278.

Schlegl, T., Seeboeck, 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 (IPMI), pp. 146–147. .

Schlegl, T., Waldstein, S.M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B.S., Langs, G., Schmidt-Erfurth, U., 2018b. Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. Ophthalmology 125, 549-558.

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, 24-30.

Schmidt-Erfurth, U., Waldstein, S.M., 2016. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Prog Retin Eye Res 50, 1-24.

Schmidt-Erfurth, U., Waldstein, S.M., Deak, G.G., Kundi, M., Simader, C., 2015. Pigment epithelial detachment followed by retinal cystoid degeneration leads to vision loss in treatment of neovascular age-related macular degeneration. Ophthalmology 122, 822-832.

Schmidt-Erfurth, U., Waldstein, S.M., Klimscha, S., Sadeghipour, A., Hu, X., Gerendas, B.S., Osborne, A., Bogunovic, H., 2018b. Prediction of Individual Disease Conversion in Early AMD using Artificial Intelligence. Investigative Ophthalmology & Visual Science submitted.

Schmitz-Valckenberg, S., Göbel, A.P., Saur, S.C., Steinberg, J.S., Thiele, S., Wojek, C., Russmann, C., Holz, F.G., for the, M.-S.G., 2016. Automated Retinal Image Analysis for Evaluation of Focal Hyperpigmentary Changes in Intermediate Age-Related Macular Degeneration. Translational Vision Science & Technology 5, 3.

Scientific-American, 2015. World Changing Ideas 2015, Scientific American.

Scott, I.U., Bressler, N.M., Bressler, S.B., Browning, D.J., Chan, C.K., Danis, R.P., Davis, M.D., Kollman, C., Qin, H., Diabetic Retinopathy Clinical Research Network Study, G., 2008. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. Retina 28, 36-40.

Seeboeck, P., Waldstein, S.M., Klimscha, S., Gerendas, B.S., Donner, R., Schlegl, T., Schmidt-Erfurth, U., Langs, G., 2016. Identifying and Categorizing Anomalies in Retinal Imaging Data. arXiv.org, arXiv:1612.00686.

Shi, F., Chen, X., Zhao, H., Zhu, W., Xiang, D., Gao, E., Sonka, M., Chen, H., 2015. Automated 3-D Retinal Layer Segmentation of Macular Optical Coherence Tomography Images With Serous Pigment Epithelial Detachments. IEEE Transactions on Medical Imaging 34, 441-452.

Sidibé, D., Sankar, S., Lemaître, G., Rastgoo, M., Massich, J., Cheung, C.Y., Tan, G.S.W., Milea, D., Lamoureux, E., Wong, T.Y., Mériaudeau, F., 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.

Silva, R., Berta, A., Larsen, M., Macfadden, W., Feller, C., Mones, J., 2018. Treat-and-Extend versus Monthly Regimen in Neovascular Age-Related Macular Degeneration: Results with Ranibizumab from the TREND Study. Ophthalmology 125, 57-65.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556

Somashekhar, S.P., Sepulveda, M.J., Puglielli, S., Norden, A.D., Shortliffe, E.H., Rohit Kumar, C., Rauthan, A., Arun Kumar, N., Patil, P., Rhee, K., Ramya, Y., 2018. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 29, 418-423.

Song, Q., Bai, J., Garvin, M.K., Sonka, M., Buatti, J.M., Wu, X., 2013. Optimal Multiple Surface Segmentation With Shape and Context Priors. IEEE Transactions on Medical Imaging 32, 376-386.

Spaide, R.F., 2018. IMPROVING THE AGE-RELATED MACULAR DEGENERATION CONSTRUCT: A New Classification System. Retina 38, 891-899.

Spaide, R.F., Fujimoto, J.G., Waheed, N.K., Sadda, S.R., Staurenghi, G., 2017. Optical coherence tomography angiography. Prog Retin Eye Res.

Srinivasan, P.P., Kim, L.A., Mettu, P.S., Cousins, S.W., Comer, G.M., Izatt, J.A., Farsiu, S., 2014. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomedical Optics Express 5, 3568-3577.

Sun, Y., Li, S., Sun, Z., 2017. Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. Journal of biomedical optics 22, 16012.

Sun, Z., Chen, H., Shi, F., Wang, L., Zhu, W., Xiang, D., Yan, C., Li, L., Chen, X., 2016. An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images. Scientific Reports 6, 21739.

Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H., 2017. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PloS one 12, e0179790.

Tang, A., Tam, R., Cadrin-Chenevert, A., Guest, W., Chong, J., Barfett, J., Chepelev, L., Cairns, R., Mitchell, J.R., Cicero, M.D., Poudrette, M.G., Jaremko, J.L., Reinhold, C., Gallix, B., Gray, B., Geis, R., Canadian Association of Radiologists Artificial Intelligence Working, G., 2018. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. Can Assoc Radiol J 69, 120-135.

Ting, D.S.W., Cheung, C.Y., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., Wong, E.Y.M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N.C., Finkelstein, E.A., Lamoureux, E.L., Wong, I.Y., Bressler, N.M., Sivaprasad, S., Varma, R., Jonas, J.B., He, M.G., Cheng, C.Y., Cheung, G.C.M., Aung, T., Hsu, W., Lee, M.L., Wong, T.Y., 2017. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 318, 2211-2223.

The Guardian, 2017. Royal Free breached UK data law in 1.6m patient deal with Google's DeepMind. https://www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act (accessed 19 July 2018).

Topol, E., 2011. The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care. Basic Books.

Torkamani, A., Andersen, K.G., Steinhubl, S.R., Topol, E.J., 2017. High-Definition Medicine. Cell 170, 828-843.

Toth, C.A., Decroos, F.C., Ying, G.S., Stinnett, S.S., Heydary, C.S., Burns, R., Maguire, M., Martin, D., Jaffe, G.J., 2015. Identification of Fluid on Optical Coherence Tomography by Treating Ophthalmologists Versus a Reading Center in the Comparison of Age-Related Macular Degeneration Treatments Trials. Retina 35, 1303-1314.

Treder, M., Lauermann, J.L., Eter, N., 2018. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. Graefes Arch Clin Exp Ophthalmol 256, 259-265.

Tufail, A., Rudisill, C., Eagan, C., Kapetanakis, V.V., Salas-Vega, S., Owen, C.G., Lee, A., Louw, V., Anderson, J., Liew, G., Bolter, L., Srinivas, S., Nittala, M., Sadda, S., Raylor, P., Rudnicka, A.R., 2017. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. Ophthalmology 124:343-351.

van Grinsven, M.J., van Ginneken, B., Hoyng, C.B., Theelen, T., Sanchez, C.I., 2016. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. IEEE Trans Med Imaging 35, 1273-1284.

van Grinsven, M.J.J.P., Buitendijk, G.H.S., Brussee, C., van Ginneken, B., Hoyng, C.B., Theelen, T., Klaver, C.C.W., Sánchez, C.I., 2015. Automatic Identification of Reticular Pseudodrusen Using Multimodal Retinal Image Analysis. Investigative Ophthalmology & Visual Science 56, 633-639.

van Grinsven, M.J.J.P., Lechanteur, Y.T.E., van de Ven, J.P.H., van Ginneken, B., Hoyng, C.B., Theelen, T., Sánchez, C.I., 2013. Automatic Drusen Quantification and Risk Assessment of Age-Related Macular Degeneration on Color Fundus Images. Investigative Ophthalmology & Visual Science 54, 3019-3027.

Varma, R., Bressler, N.M., Doan, Q.V., Danese, M., Dolan, C.M., Lee, A., Turpcu, A., 2015. Visual Impairment and Blindness Avoided with Ranibizumab in Hispanic and Non-Hispanic Whites with Diabetic Macular Edema in the United States. Ophthalmology 122, 982-989.

Veiga, D., Martins, N., Ferreira, M., Monteiro, J., 2017. Automatic microaneurysm detection using laws texture masks and support vector machines. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 0, 1-12.

Venhuizen, F.G., Ginneken, B.v., Liefers, B., Asten, F.v., Schreur, V., Fauser, S., Hoyng, C., Theelen, T., Sánchez, C.I., 2018. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomedical Optics Express 9, 1545-1569.

Venhuizen, F.G., van Ginneken, B., van Asten, F., van Grinsven, M., Fauser, S., Hoyng, C.B., Theelen, T., Sanchez, C.I., 2017. Automated Staging of Age-Related Macular Degeneration Using Optical Coherence Tomography. Invest Ophthalmol Vis Sci 58, 2318-2328.

Verghese, A., Shah, N.H., Harrington, R.A., 2018. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA 319, 19-20.

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. Invest Ophthalmol Vis Sci 58, 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 Trans Med Imaging 36, 1773-1783.

Waldstein, S.M., Gerendas, B.S., Montuoro, A., Simader, C., Schmidt-Erfurth, U., 2015. Quantitative comparison of macular segmentation performance using identical retinal regions across multiple spectral-domain optical coherence tomography instruments. The British journal of ophthalmology 99, 794-800.

Waldstein, S.M., Montuoro, A., Podkowinski, D., Philip, A.M., Gerendas, B.S., Bogunovic, H., Schmidt-Erfurth, U., 2017. Evaluating the impact of vitreomacular adhesion on anti-VEGF therapy for retinal vein occlusion using machine learning. Sci Rep 7, 2928.

Waldstein, S.M., Philip, A.M., Leitner, R., Simader, C., Langs, G., Gerendas, B.S., Schmidt-Erfurth, U., 2016. Correlation of 3-Dimensionally Quantified Intraretinal and Subretinal Fluid With Visual Acuity in Neovascular Age-Related Macular Degeneration. JAMA Ophthalmol 134, 182-190.

Walter, T., Massin, P., Erginay, A., Ordonez, R., Jeulin, C., Klein, J.-C., 2007. Automatic detection of microaneurysms in color fundus images. Medical Image Analysis 11, 555-566.

Wang, S., Tang, H.L., turk, L.I.A., Hu, Y., Sanei, S., Saleh, G.M., Peto, T., 2017a. Localizing Microaneurysms in Fundus Images Through Singular Spectrum Analysis. IEEE Transactions on Biomedical Engineering 64, 990-1002.

Wang, S.K., Callaway, N.F., Wallenstein, M.B., Henderson, M.T., Leng, T., Moshfeghi, D.M., 2015. SUNDROP: six years of screening for retinopathy of prematurity with telemedicine. Canadian journal of ophthalmology. Journal canadien d'ophtalmologie 50, 101-106.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017b. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv:1705.02315

Wang, Z., Camino, A., Hagag, A.M., Wang, J., Weleber, R.G., Yang, P., Pennesi, M.E., Huang, D., Li, D., Jia, Y., 2018. Automated detection of preserved photoreceptor on optical coherence tomography in choroideremia based on machine learning. Journal of Biophotonics.

Welikala, R.A., Foster, P.J., Whincup, P.H., Rudnicka, A.R., Owen, C.G., Strachan, D.P., Barman, S.A., 2017. Automated arteriole and venule classification using deep learning for retinal images from the UK Biobank cohort. Computers in Biology and Medicine 90, 23-32.

Wells, J.A., Glassman, A.R., Jampol, L.M., Aiello, L., Antoszyk, A., Arnold-Bush, B., Baker, C., Bressler, N., Browning, D., Elman, M., Ferris, F., Friedman, C., Pieramici, D., Sun, J., Beck, R., 2015. Aflibercept, Bevacizumab, or Ranibizumab for Diabetic Macular Edema. New England Journal of Medicine 372, 1193-1203.

Yan, Q., Ding, Y., Liu, Y., Sun, T., Fritsche, L.G., Clemons, T., Ratnapriya, R., Klein, M.L., Cook, R.J., Liu, Y., Fan, R., Wei, L., Abecasis, G.R., Swaroop, A., Chew, E.Y., Group, A.R., Weeks, D.E., Chen, W., 2018. Genome-wide analysis of disease progression in age-related macular degeneration. Hum Mol Genet 27, 929-940.

Yu, F., Sun, J., Li, A., Cheng, J., Wan, C., Liu, J., 2017. Image quality classification for DR screening using deep learning, 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 664-667.

Yu, V.L., Fagan, L.M., Wraith, S.M., Clancey, W.J., Scott, A.C., Hannigan, J., Blum, R.L., Buchanan, B.G., Cohen, S.N., 1979. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. JAMA 242, 1279-1282.

Zadeh, S.G., Wintergerst, M., Wiens, V., Thiele, S., Holz, F., Finger, R., Schultz, T., 2017. CNNs Enable Accurate and Fast Segmentation of Drusen in Optical Coherence Tomography. Proceedings of Deep Learning in Medical Image Analysis (DLMIA).

Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks, European Conference on Computer Vision Springer, pp. 818-833.

Zhang, L., Sonka, M., Folk, J.C., Russell, S.R., Abramoff, M.D., 2014. Quantifying disrupted outer retinal-subretinal layer in SD-OCT images in choroidal neovascularization. Invest Ophthalmol Vis Sci 55, 2329-2335.

Zheng, Y., Xiao, R., Wang, Y., Gee, J.C., 2013. A generative model for OCT retinal layer segmentation by integrating graph-based multi-surface searching and image registration. Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 16, 428-435.

Ziemssen, F., Bertelmann, T., Hufenbach, U., Scheffler, M., Liakopoulos, S., Schmitz-Valckenberg, S., 2016. [Delayed treatment initiation of more than 2 weeks. Relevance for possible gain of visual acuity after anti-VEGF therapy under real life conditions (interim analysis of the prospective OCEAN study)]. Ophthalmologe 113, 143-151.

Legend

Figures

Figure 1: Spectral-domain OCT cross-section (B-scan) of the macula, showing a wealth of detailed pathognomonic features with µm resolution over a three-dimensional volume of macular tissue. High-resolution imaging allows to quantify established clinical biomarkers as well as sub-clinical features not detectable by ophthalmic examination such as hyperreflective foci, photoreceptor alteration and feature quantification.

Figure 2: Diagnostic imaging is currently the highest and most efficient application of AI-based analyses and will likely further expand as imaging modalities become advance and multi-modal. (Jiang et al., 2017)

Figure 3: Illustration of the performance advantage of the deep learning models over the classic machine learning ones. The difference becomes clear as the amount of training data keeps increasing. Unlike the performance of classic machine learning methods which tend to saturate with the amount of data, the performance of deep learning keeps growing.

Figure 4: Illustration of the principal difference between the classic machine learning models and the deep learning ones. Instead of learning to classify an input image from hand-engineered features deep learning models are learning to both extract features and classify from the input directly, hence allowing for fully "end-to-end" learning.

Figure 5: Machine learning approaches focus on the learning feature of artificial intelligence. These and in particular deep learning methods have been successfully applied in different use-cases of retina imaging: classification, segmentation and prediction based on OCT and CFP. Image processing methods do not belong the field of AI but they have achieved comparable results, especially in CFP segmentation.

Figure 6: Recent advances in the segmentation of CFP and OCT show comparable performance between human graders and machine learning algorithms. In this context, for many segmentation targets the difference between two graders is comparable to each of them relative to the automated segmentation. [Modified after Bogunovic et al., 2017b)

Figure 7: Prediction of individual treatment requirements during as-needed anti-VEGF therapy. AI models are trained to utilize data from the common loading dose to predict individual treatment needs during long-term follow-up. The currently available methods achieve an accuracy of about 70%-80%, which is superior to human performance. (Schmidt-Erfurth et al., 2018a)

Figure 8: Extraction of imaging biomarkers to allow prediction of visual acuity outcomes. The OCT images acquired during the loading phase were analysed automatically by deep learning and graph cut tools, yielding spatially resolved measurements of intraretinal fluid, subretinal fluid, pigment epithelial detachment as well as retinal thickness, among other markers. The obtained variables were fed into the AI modelling database. (Niu et al., 2016)

Figure 9: Prediction of GA growth over time using machine learning. An example of a GA lesion which grows continuously over time is shown. The AI algorithm provides a probability map for future growth (right column). The accuracy of the algorithm is illustrated in the center column. This study was successful in forecasting the future development of GA lesions, although the results were not compared against a benchmark (i.e. assuming linear growth in all directions by the known growth rates). (Bogunovic et al., 2017a)

Figure 10: Prediction of drusen growth and drusen regression using AI tools. Three example patients are provided (rows 1-3). All patients experienced drusen regression, an important hallmark of AMD progression at Year 1 compared to baseline (“Gold standard”). The AI method achieved an 80% performance in predicting the future time and location of drusen regression (“Prediction”). [Modified after de Sisternes et al., 2014]

Figure 11: AI to predict the risk of AMD conversion on a patient level. From a set of quantitative features extracted from drusen, the progression risk of advanced AMD onset could be successfully determined. The AI system assigns each patient with a hazard ratio; the allocation was highly accurate as shown in the Kaplan Meier Plot on the right. [Modified after Schmidt-Erfurth et al., 2018b]

Figure 12: AI may not only predict and differentiate a priori the development of CNV and/or GA in AMD eyes (left), but also provides insight into the pathophysiologic fingerprint of AMD biomarkers (right). While the development of CNV is almost exclusively driven by drusen-associated changes, the risk of GA is closely related to (atrophic) changes in the outer neurosensory retina, hyperreflective foci and age.

Figure 13: Interpretation of a deep learning model's output for the detection of diabetic retinopathy. Color-coded map obtained from the model is overlaid on a fundus image highlighting pathologic regions on which the decision was based. [Adapted from from Gargeya and Leng, 2017]

Figure 14: Prediction of age from fundus image (left) using a deep learning model and the corresponding heatmap overlaid in green (right) indicating the areas of the fundus that the neural network model is relying on to make the prediction. [Adapted from Poplin et al., 2018]

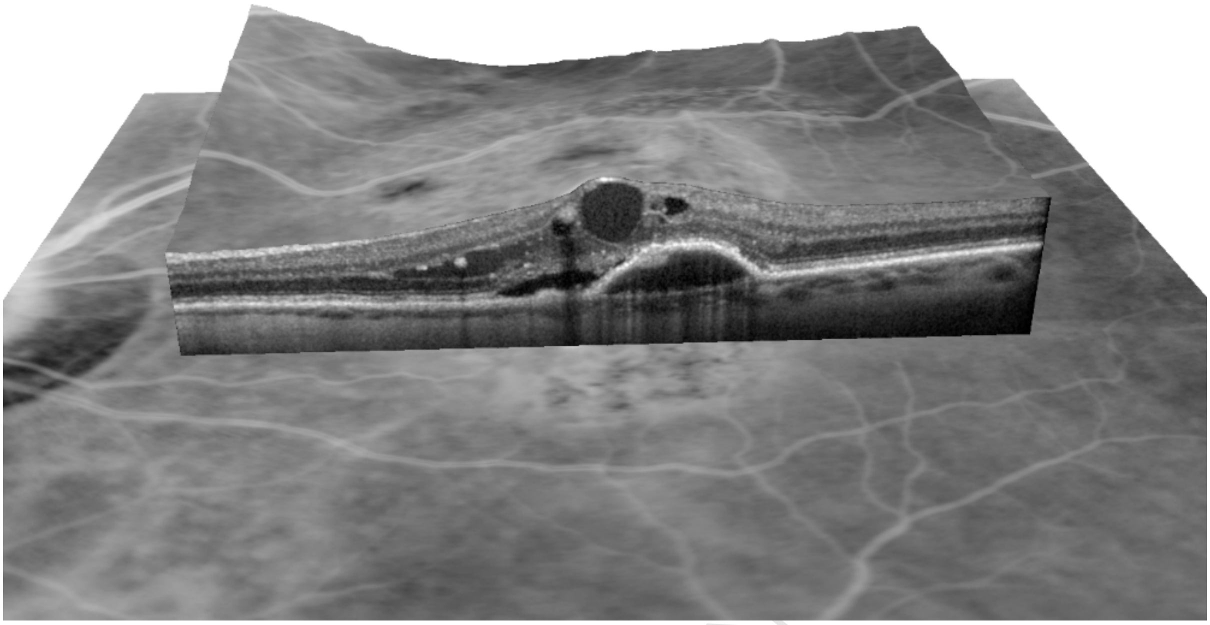
Figure 15: Examples for structures in color fundus photography. On the left all available anatomic landmarks are used for orientation: fovea, macula, blood vessels, optic nerv head, center of optic nerve head.(Molina-Casado, Carmona et al. 2017) On the right: orientation from left scan allows subsumption of relevance of pathologic structures; microaneurysms and hemorrhage are visible in this image; the corners show exemplary image patches for variability of microaneurysms that algorithms are comparing with.(Moccia et al., 2018)

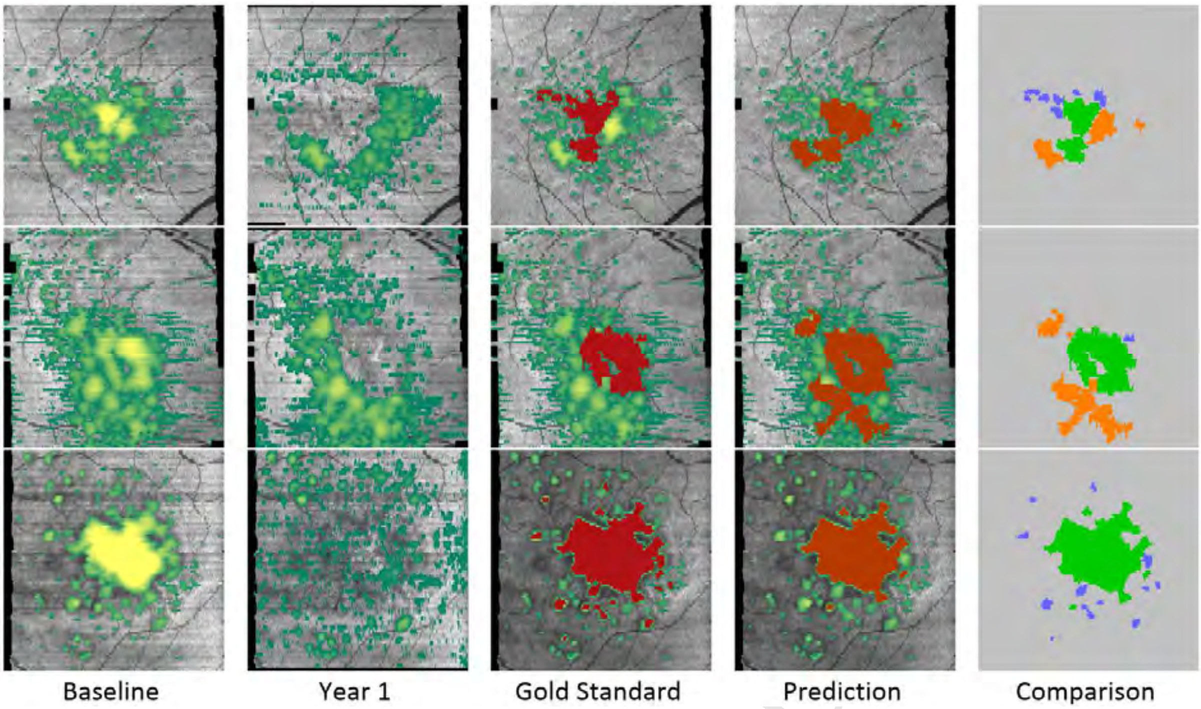
Figure 16: Fully automated quantification of intra- and subretinal macular fluid by deep learning. This method was validated in 1,200 eyes, 3 diseases and 2 OCT devices and achieved a clinically applicable accuracy of 90%-96%. Upper row: OCT b-scans; middle row: ground truth = manual annotation by human grader; lower row: automated result for intra- and subretinal fluid quantification. (Schlegl et al., 2017)

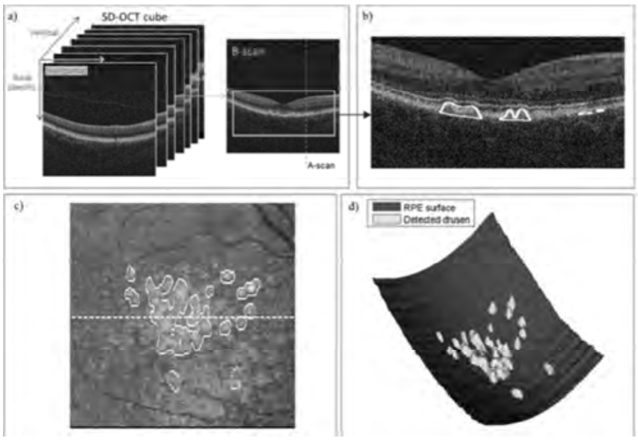
Figure 17: This figure shows a schematic overvue of the iDx-DR algorithm performance for Diabetic Retinopathy (DR) Screening: first, a quality assessment decides if the image can be used for analysis or if there are dark areas, areas that are not sharp enough or a generally wrong location of the image; second a deep learning algorithm using convolutional neuronal networks screens the image for clinical biomarkers (e.g. microaneurysms, hemorrhages, exsudates, etc.); as a last step the disease assessment with inputs from both the clinical biomarker assessment as well as the use of an anatomical location definition is performed for clinical decision and classification into no DR, moderate DR or vision-threatening DR. This device has received approval for clinical use by the food and drug administration (FDA) in April 2018.

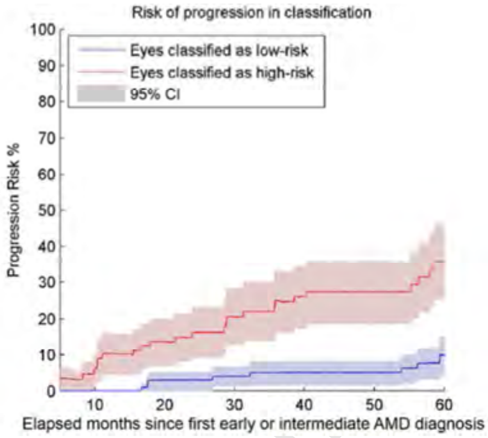
Figure 18: Venhuizen et al. assess risk stages in patients with age-related macular degeneration (AMD). Examples of b-scans showing the different severity stages of AMD as defined by a central reading center: (a) No AMD, (b) early AMD, (c) intermediate AMD, (d) advanced AMD with GA, and (e) advanced AMD with CNV. (Venhuizen et al., 2017)

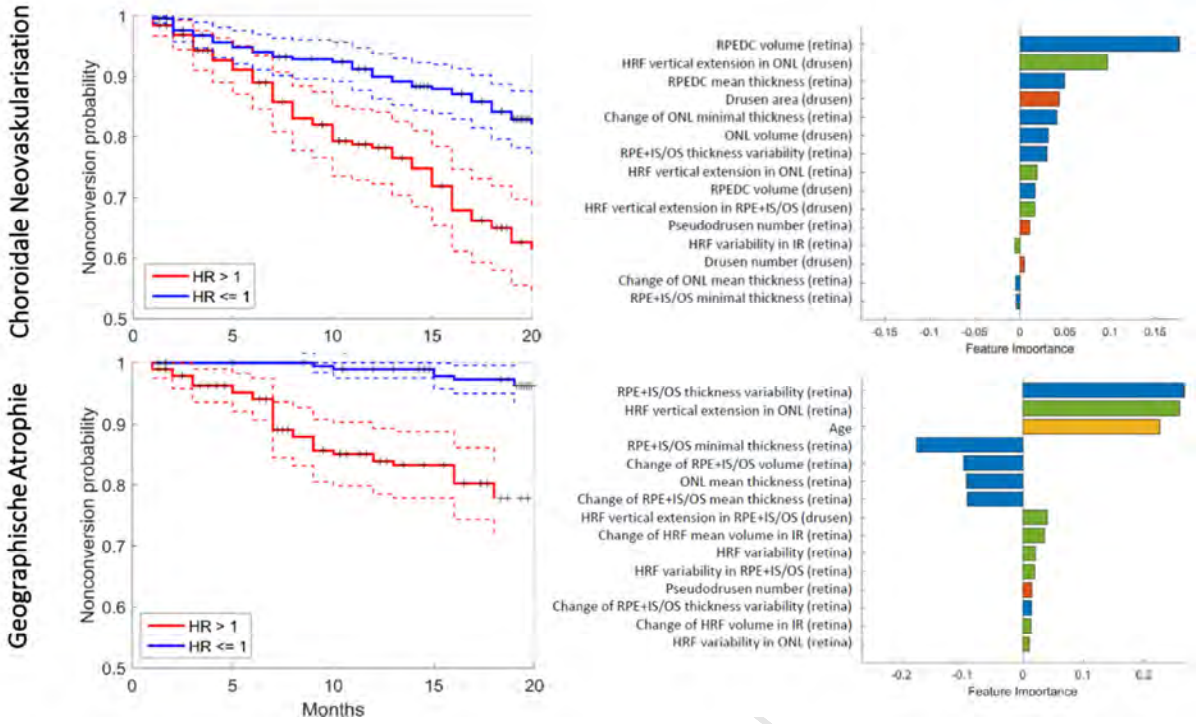
Figure 19: Research data is relevant insight which should systematically be shared with the entire academic community in a structured way pertaining particularly to publicly funded research to increase the available knowledge. (Hahnel, 2015)

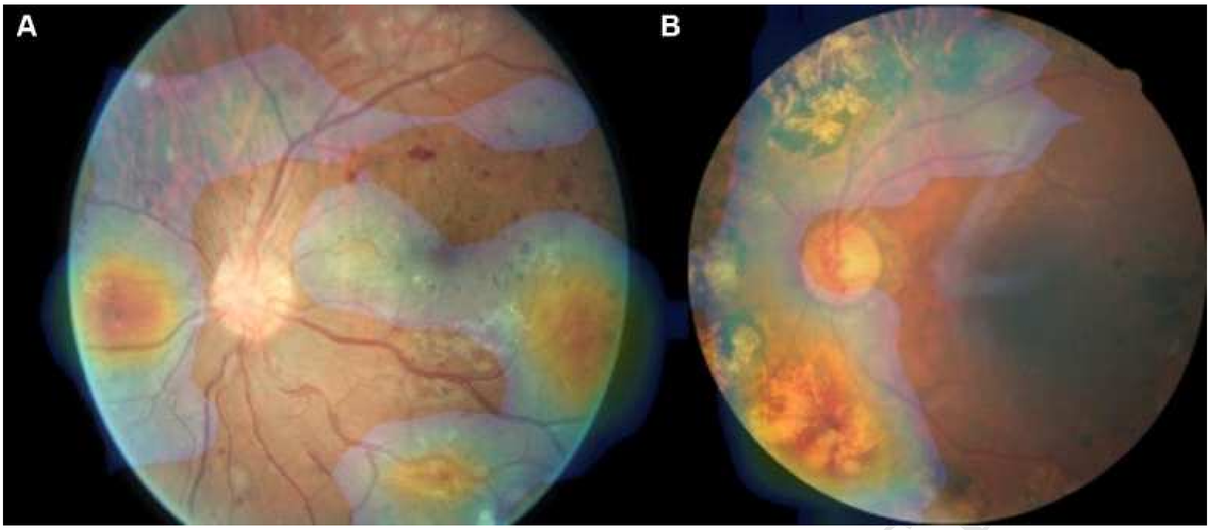


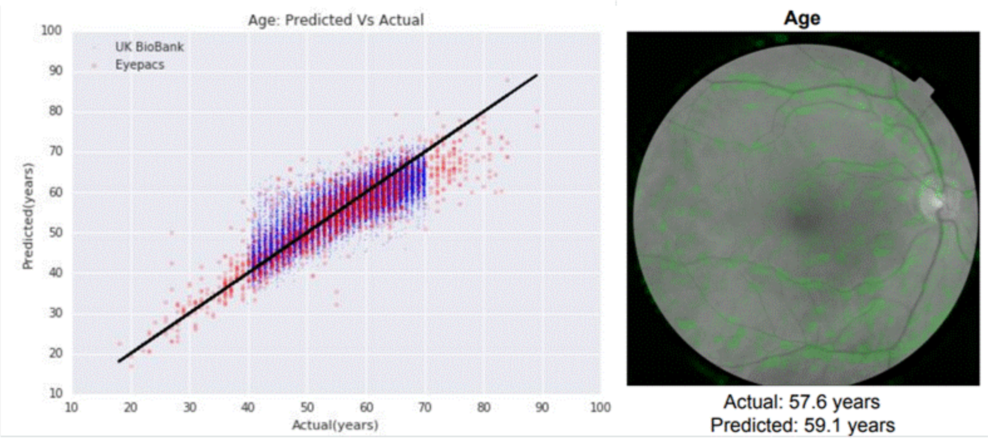

Baseline
Year 1
Gold Standard
Prediction
Comparison

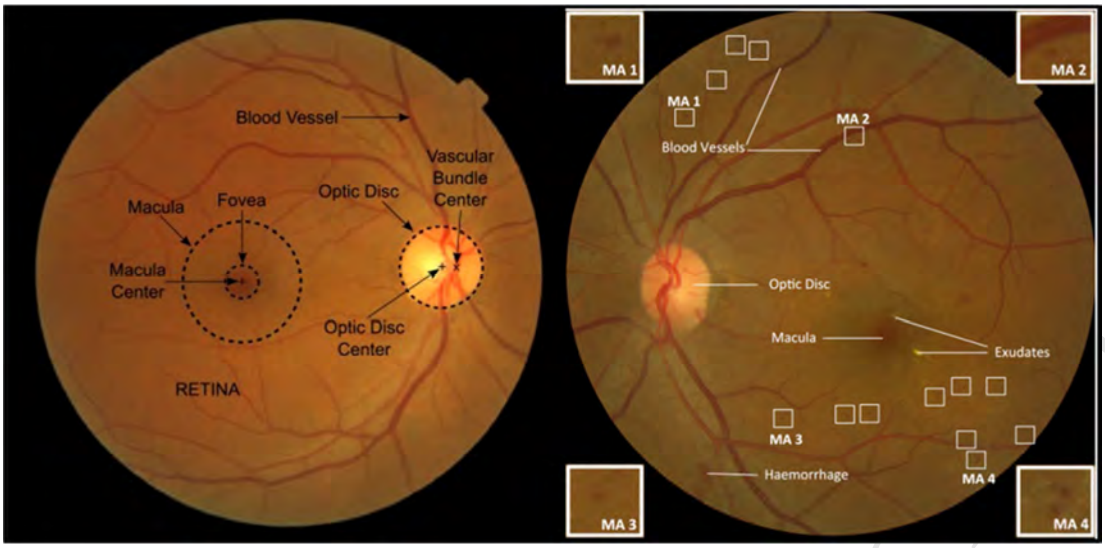

a)
SD-OCT cube
Vertical
Axial (depth)
Horizontal
B-scan
A-scan
b)
c)
d)
RPE surface
Detected drusen

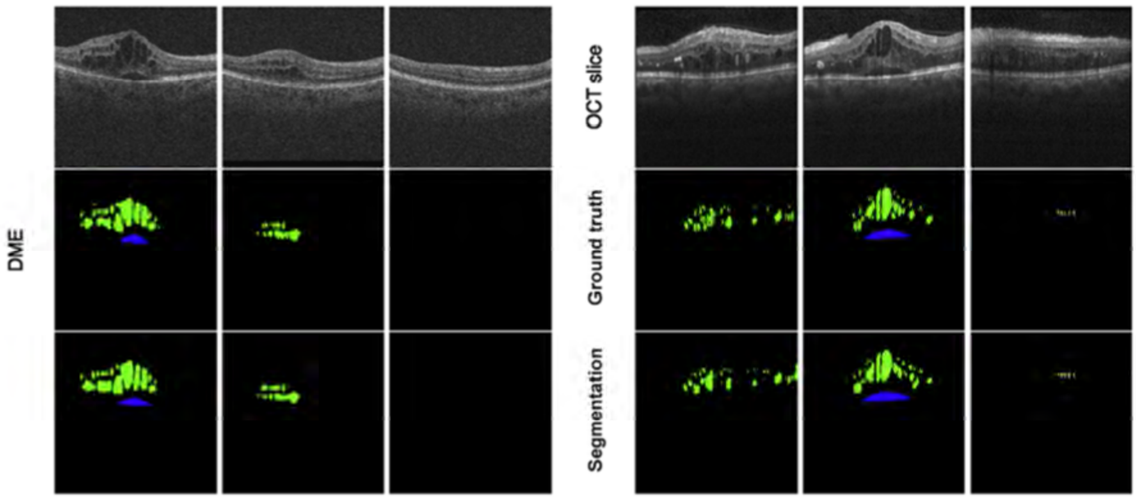

Risk of progression in classification
Eyes classified as low-risk
Eyes classified as high-risk
95% CI
Progression Risk %
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
Elapsed months since first early or intermediate AMD diagnosis

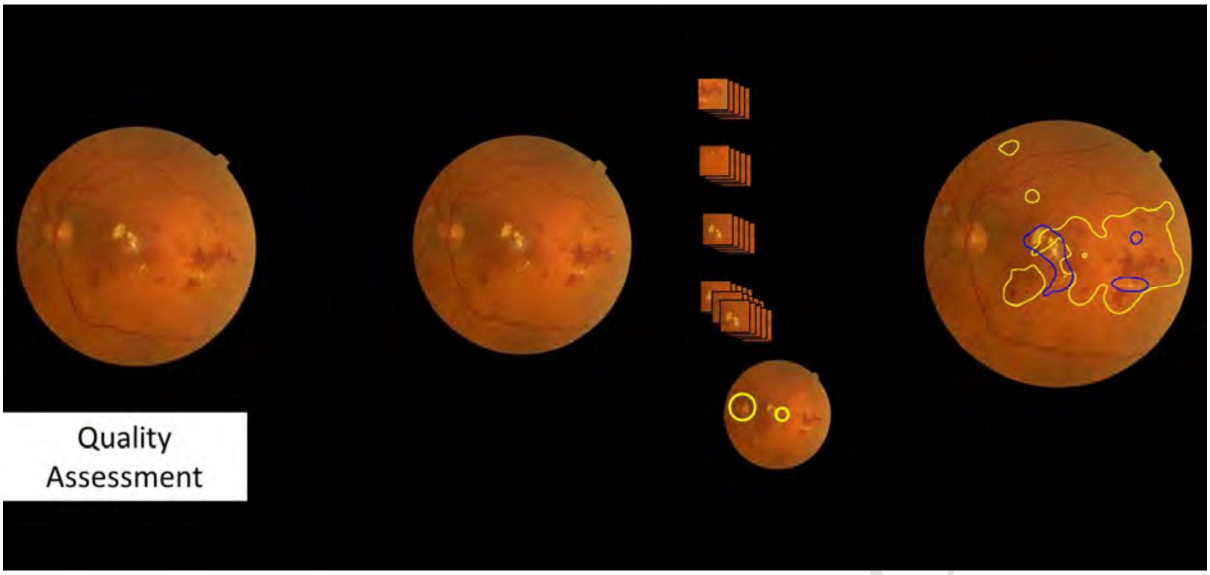

Choroidale Neovaskularisation
Geographische Atrophie
Nonconversion probability
1
0.9
0.8
0.7
0.6
0.5
0
5
10
15
20
Months
HR > 1
HR <= 1
RPEDC volume (retina)
HRF vertical extension in ONL (drusen)
RPEDC mean thickness (retina)
Drusen area (drusen)
Change of ONL minimal thickness (retina)
ONL volume (drusen)
RPE+IS/OS thickness variability (retina)
HRF vertical extension in ONL (retina)
RPEDC volume (drusen)
HRF vertical extension in RPE+IS/OS (drusen)
Pseudodrusen number (retina)
HRF variability in IR (retina)
Drusen number (drusen)
Change of ONL mean thickness (retina)
RPE+IS/OS minimal thickness (retina)
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
Feature Importance
RPE+IS/OS thickness variability (retina)
HRF vertical extension in ONL (retina)
Age
RPE+IS/OS minimal thickness (retina)
Change of RPE+IS/OS volume (retina)
ONL mean thickness (retina)
Change of RPE+IS/OS mean thickness (retina)
HRF vertical extension in RPE+IS/OS (drusen)
Change of HRF mean volume in IR (retina)
HRF variability (retina)
HRF variability in RPE+IS/OS (retina)
Pseudodrusen number (retina)
Change of RPE+IS/OS thickness variability (retina)
Change of HRF volume in IR (retina)
HRF variability in ONL (retina)
-0.2
-0.1
0
0.1
0.2
Feature Importance

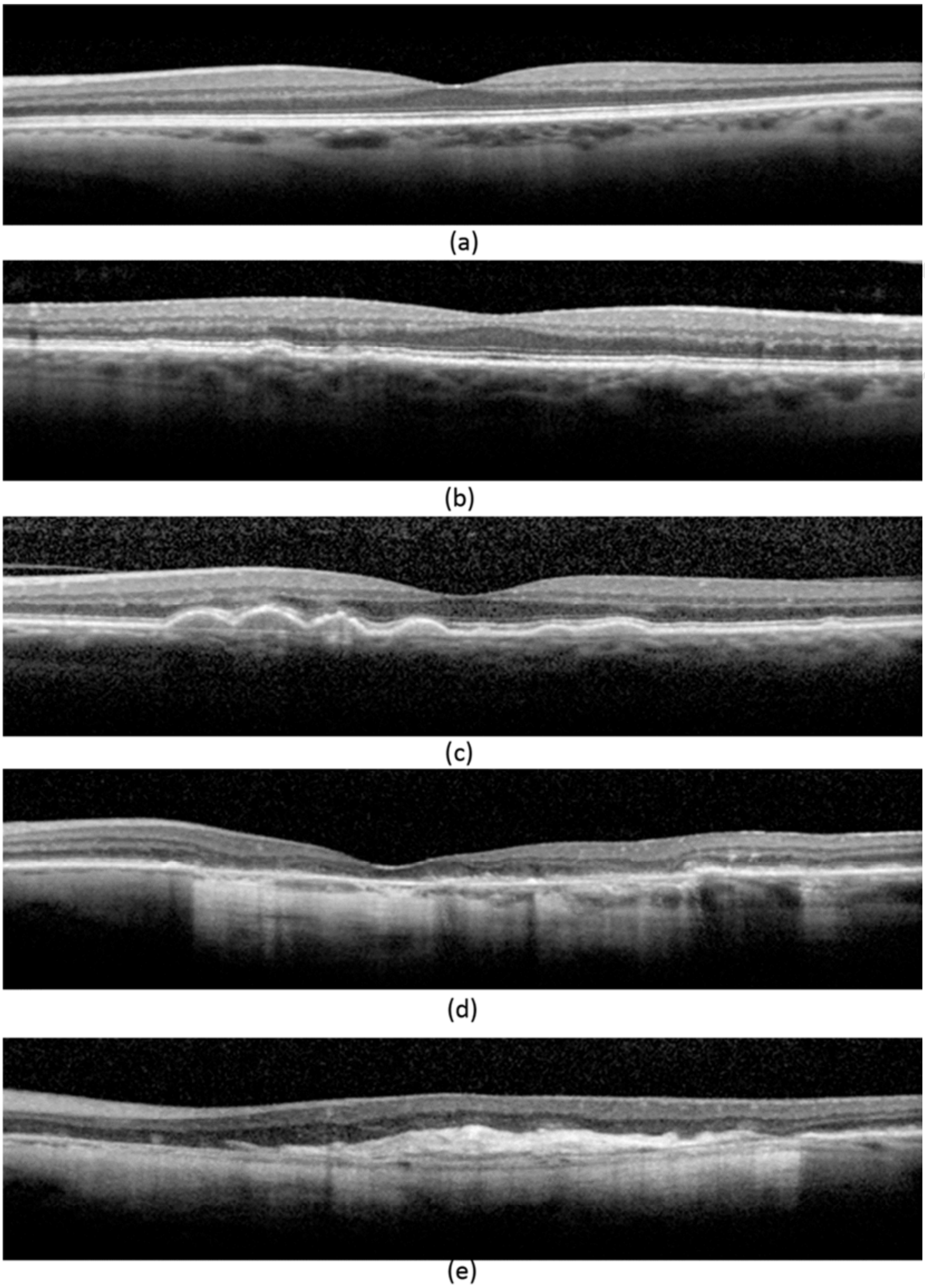

A
B

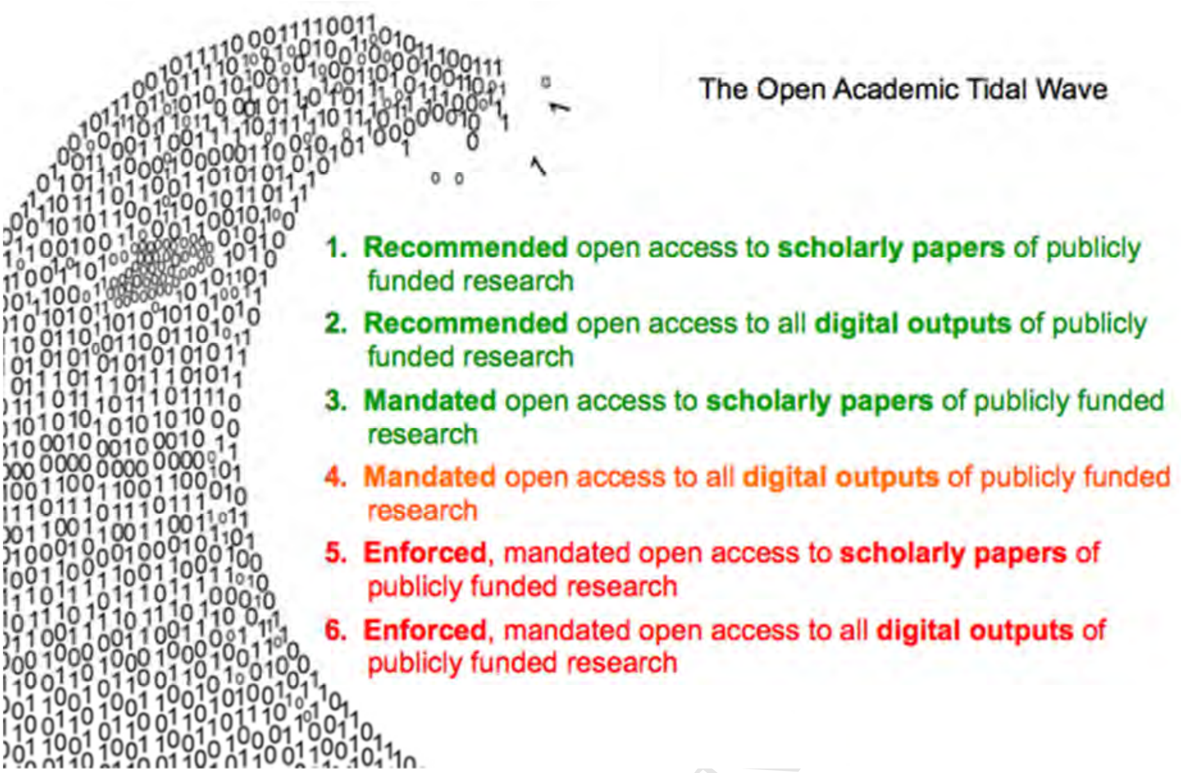

Age: Predicted Vs Actual
UK BioBank
Eyepacs
Predicted(years)
Actual(years)
10
20
30
40
50
60
70
80
90
100
Age
Actual: 57.6 years
Predicted: 59.1 years


Blood Vessel
Vascular
Bundle
Center
Optic Disc
Macula
Fovea
Macula
Center
Optic Disc
Center
RETINA
MA 1
MA 2
MA 1
MA 2
Blood Vessels
Optic Disc
Macula
Exudates
MA 3
MA 4
Haemorrhage
MA 3
MA 4


DME
OCT slice
Ground truth
Segmentation

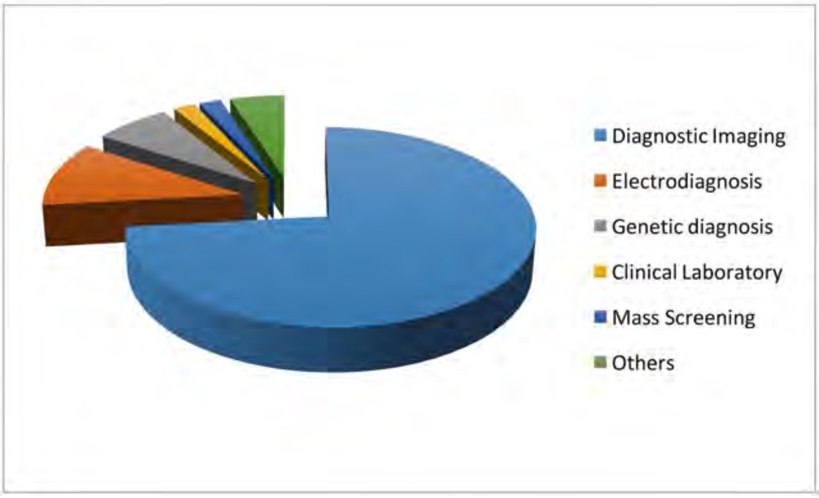

Quality
Assessment

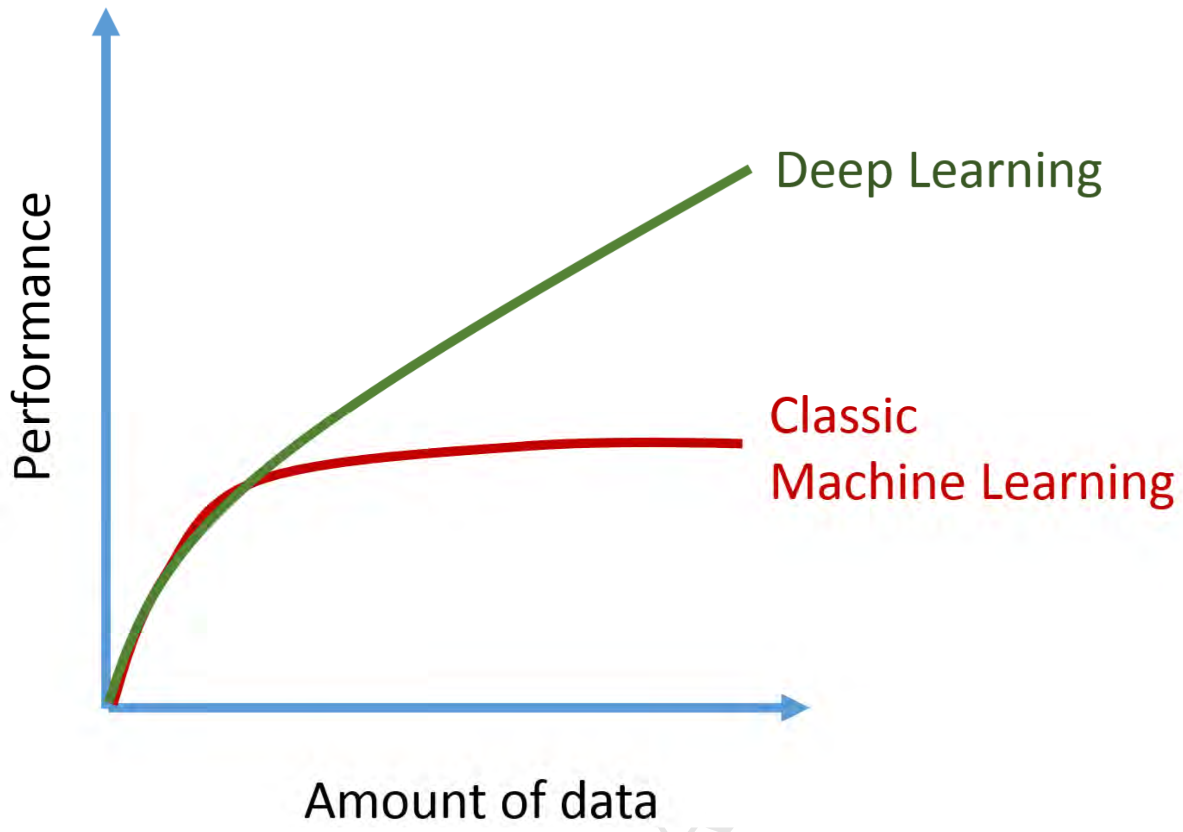

(a)
(b)
(c)
(d)
(e)

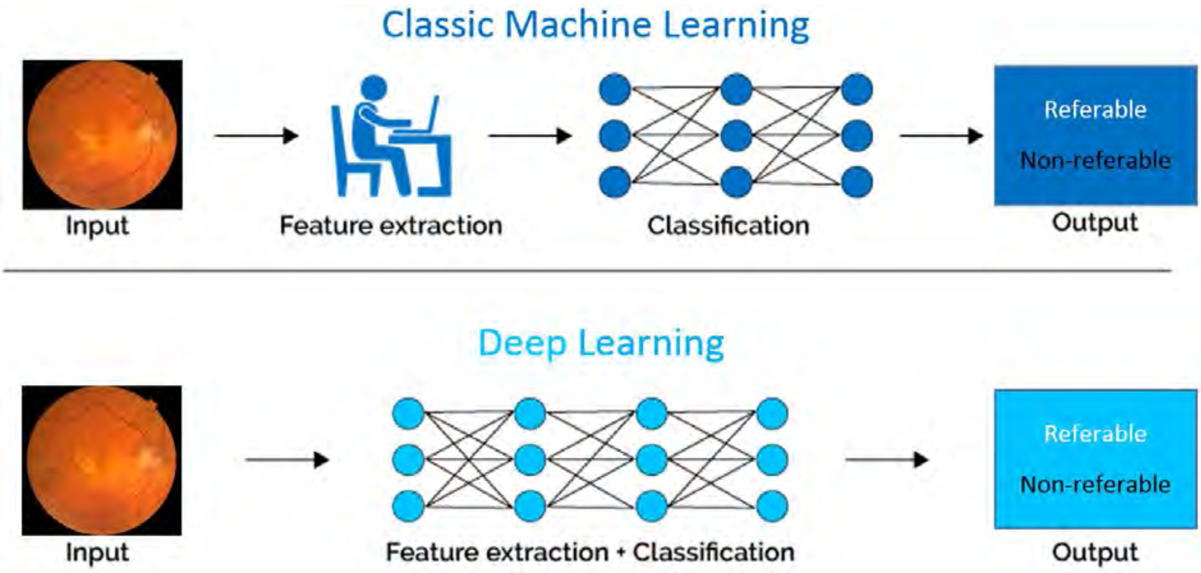


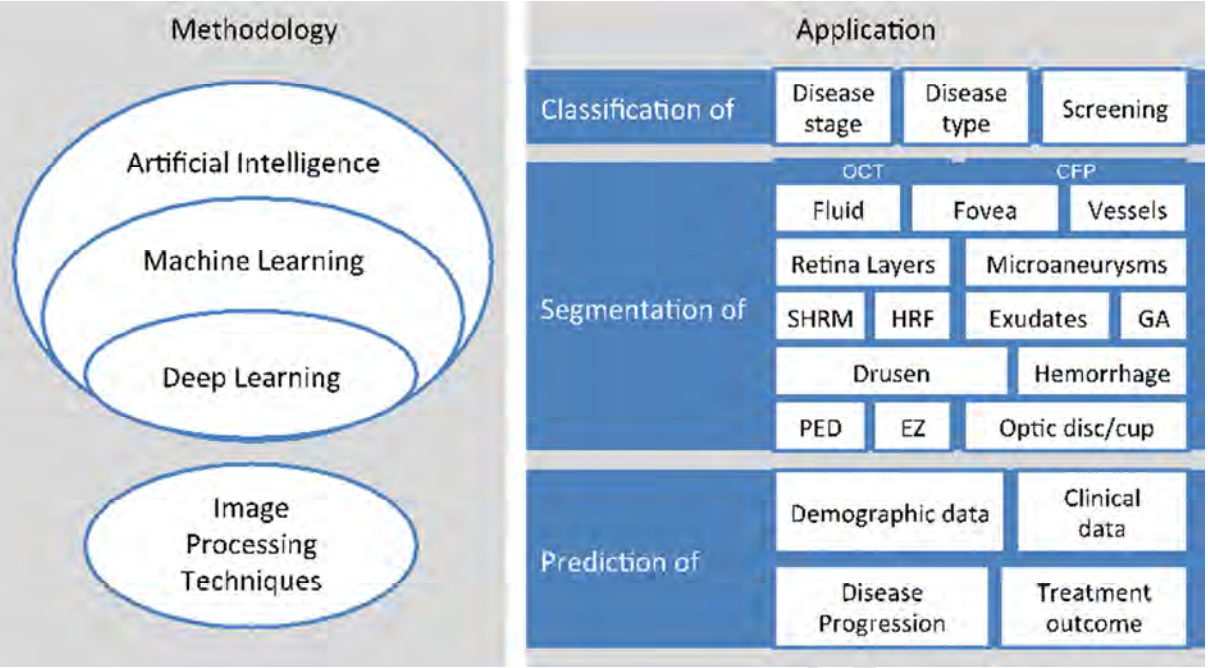
The Open Academic Tidal Wave

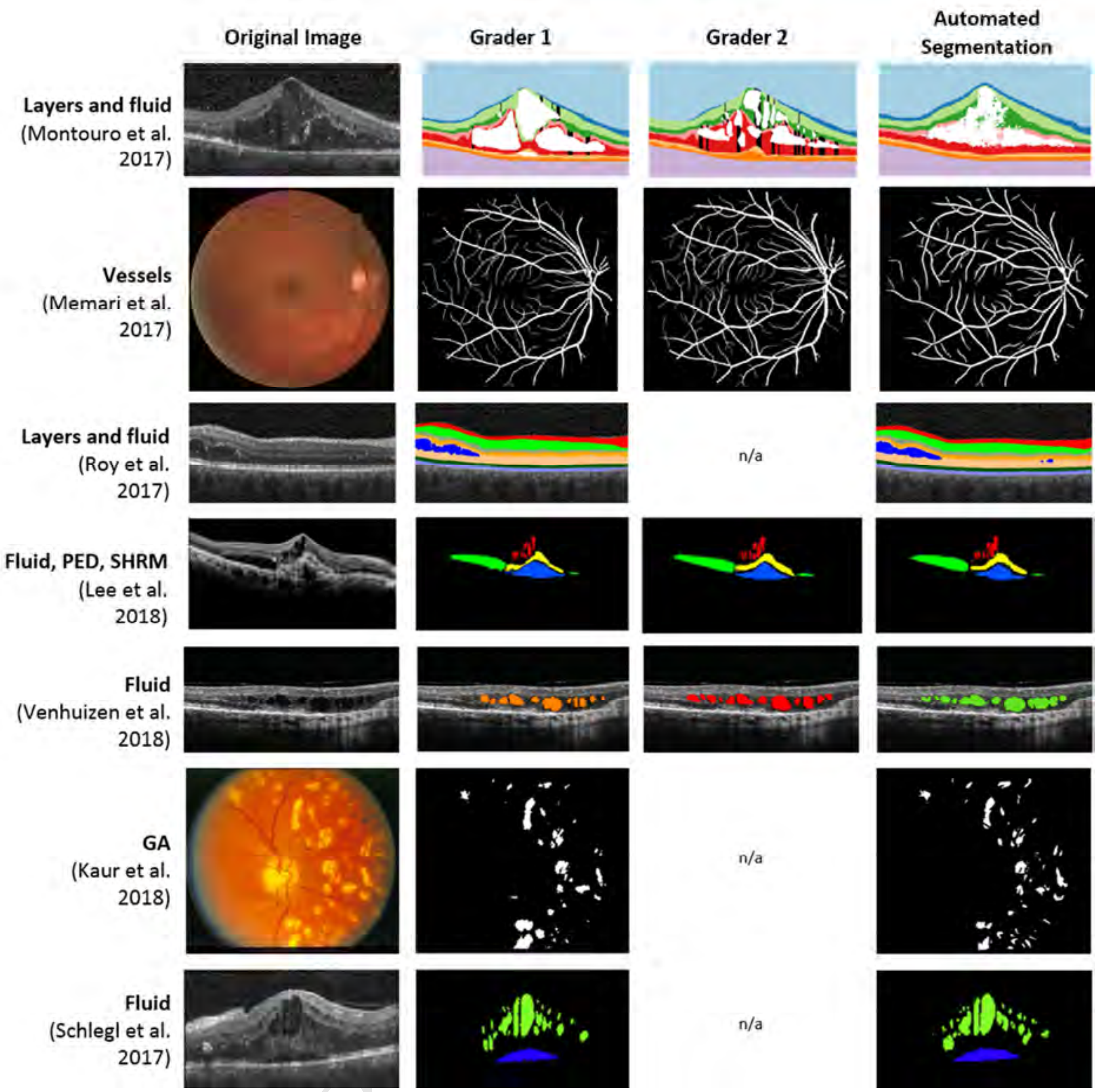
1. **Recommended** open access to **scholarly papers** of publicly funded research
2. **Recommended** open access to all **digital outputs** of publicly funded research
3. **Mandated** open access to **scholarly papers** of publicly funded research
4. **Mandated** open access to all **digital outputs** of publicly funded research
5. **Enforced**, mandated open access to **scholarly papers** of publicly funded research
6. **Enforced**, mandated open access to all **digital outputs** of publicly funded research

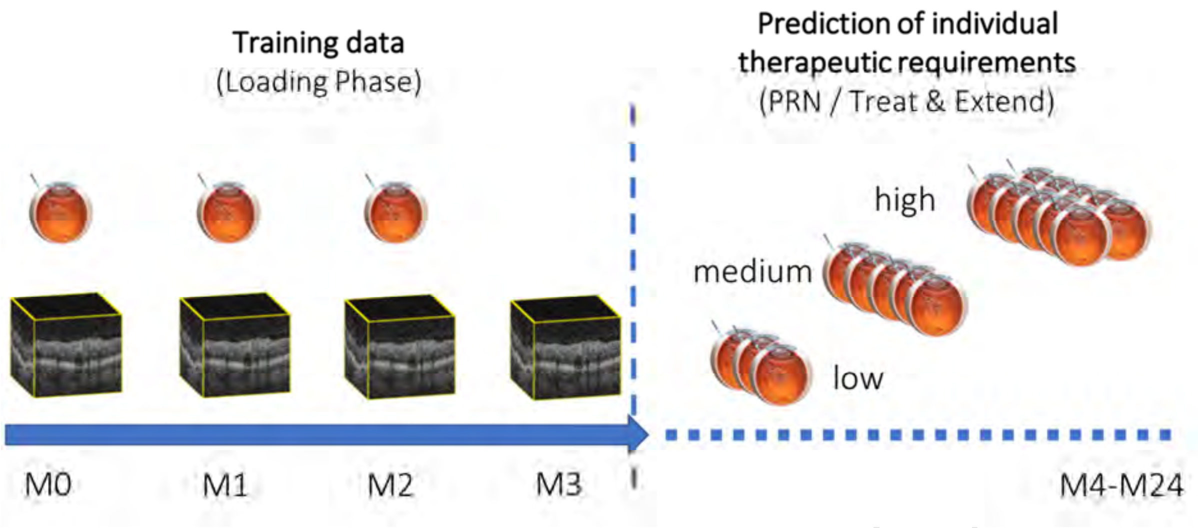

Diagnostic Imaging
Electrodiagnosis
Genetic diagnosis
Clinical Laboratory
Mass Screening
Others

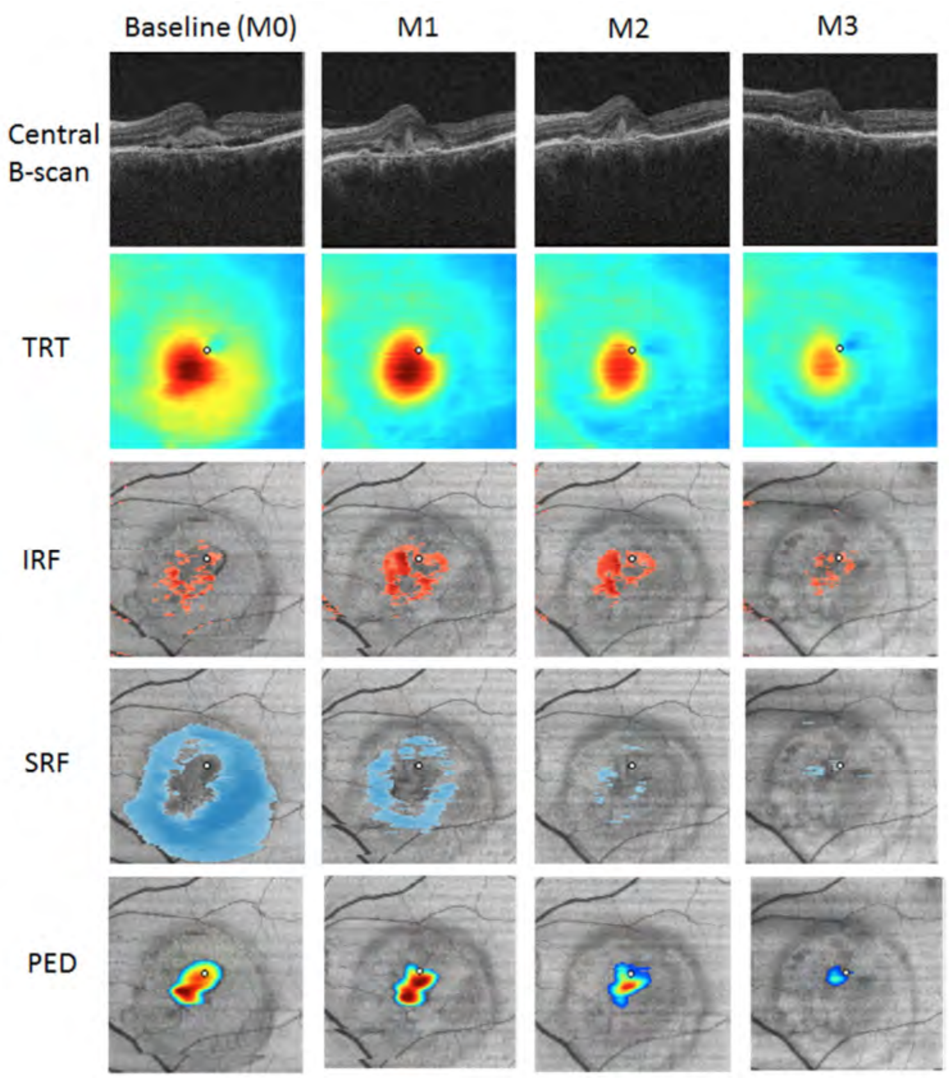

Performance
Deep Learning
Classic
Machine Learning
Amount of data

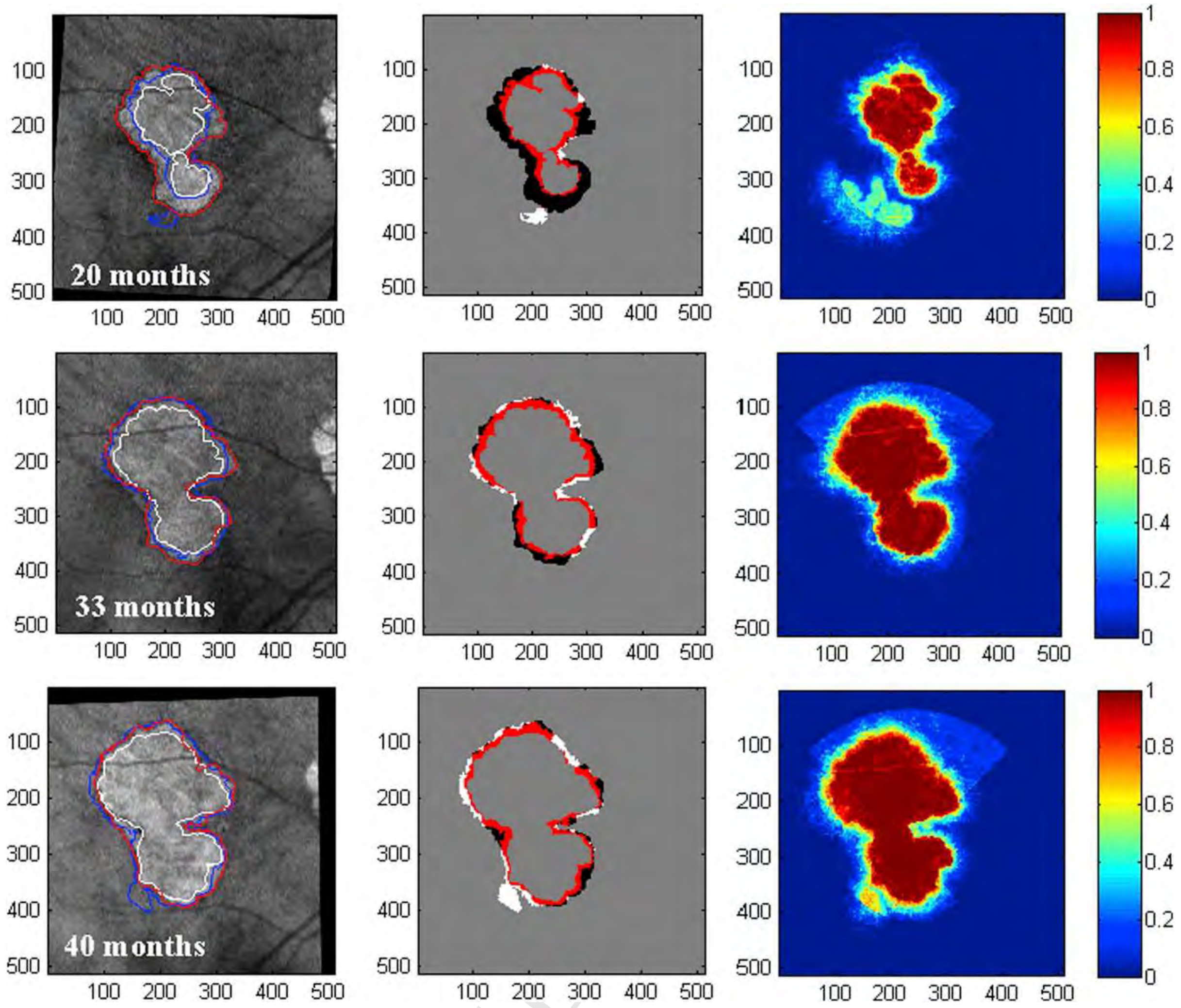

Classic Machine Learning
Input
Feature extraction
Classification
Referable
Non-referable
Output
Deep Learning
Input
Feature extraction + Classification
Referable
Non-referable
Output


Methodology
Artificial Intelligence
Machine Learning
Deep Learning
Image Processing Techniques
Application
Classification of
Disease stage
Disease type
Screening
Segmentation of
OCT
CFP
Fluid
Fovea
Vessels
Retina Layers
Microaneurysms
SHRM
HRF
Exudates
GA
Drusen
Hemorrhage
PED
EZ
Optic disc/cup
Prediction of
Demographic data
Clinical data
Disease Progression
Treatment outcome


Original Image
Grader 1
Grader 2
Automated Segmentation
Layers and fluid (Montouro et al. 2017)
Vessels (Memari et al. 2017)
Layers and fluid (Roy et al. 2017)
n/a
Fluid, PED, SHRM (Lee et al. 2018)
Fluid (Venhuizen et al. 2018)
GA (Kaur et al. 2018)
n/a
Fluid (Schlegl et al. 2017)
n/a


Training data
(Loading Phase)
Prediction of individual
therapeutic requirements
(PRN / Treat & Extend)
high
medium
low
M0
M1
M2
M3
M4-M24


Baseline (M0)
M1
M2
M3
Central B-scan
TRT
IRF
SRF
PED


20 months
33 months
40 months
100
200
300
400
500
0
0.2
0.4
0.6
0.8
1