Graefe's Archive for Clinical and Experimental Ophthalmology (2022) 260:2217–2230
https://doi.org/10.1007/s00417-022-05565-1

RETINAL DISORDERS

Clinically applicable deep learning-based decision aids for treatment of neovascular AMD

Matthias Gutfleisch[1,2] · Oliver Ester[3] · Sökmen Aydin[3] · Martin Quassowski[3] · Georg Spital[1,2] · Albrecht Lommatzsch[1,2,4,5] · Kai Rothaus[1,2] · Adam Michael Dubis[6] · Daniel Pauleikhoff[1,2,4,5]

Received: 1 June 2021 / Revised: 6 January 2022 / Accepted: 11 January 2022 / Published online: 22 January 2022

Abstract

Purpose Anti-vascular endothelial growth factor (Anti-VEGF) therapy is currently seen as the standard for treatment of neovascular AMD (nAMD). However, while treatments are highly effective, decisions for initial treatment and retreatment are often challenging for non-retina specialists. The purpose of this study is to develop convolutional neural networks (CNN) that can differentiate treatment indicated presentations of nAMD for referral to treatment centre based solely on SD-OCT. This provides the basis for developing an applicable medical decision support system subsequently.

Methods SD-OCT volumes of a consecutive real-life cohort of 1503 nAMD patients were analysed and two experiments were carried out. To differentiate between no treatment class vs. initial treatment nAMD class and stabilised nAMD vs. active nAMD, two novel CNNs, based on SD-OCT volume scans, were developed and tested for robustness and performance. In a step towards explainable artificial intelligence (AI), saliency maps of the SD-OCT volume scans of 24 initial indication decisions with a predicted probability of > 97.5% were analysed (score 0–2 in respect to staining intensity). An AI benchmark against retina specialists was performed.

Results At the first experiment, the area under curve (AUC) of the receiver-operating characteristic (ROC) for the differentiation of patients for the initial analysis was 0.927 (standard deviation (SD): 0.018), for the second experiment (retreatment analysis) 0.865 (SD: 0.027). The results were robust to downsampling (¼ of the original resolution) and cross-validation (tenfold). In addition, there was a high correlation between the AI analysis and expert opinion in a sample of 102 cases for differentiation of patients needing treatment ($\kappa = 0.824$). On saliency maps, the relevant structures for individual initial indication decisions were the retina/vitreous interface, subretinal space, intraretinal cysts, subretinal pigment epithelium space, and the choroid.

Conclusion The developed AI algorithms can define and differentiate presentations of AMD, which should be referred for treatment or retreatment with anti-VEGF therapy. This may support non-retina specialists to interpret SD-OCT on expert opinion level. The individual decision of the algorithm can be supervised by saliency maps.

Keywords Neovascular age-related macular degeneration (nAMD) · Anti-VEGF therapy · Artificial intelligence · Deep learning network · Convolutional neural network · Treatment algorithms

✉ Matthias Gutfleisch
Matthias.Gutfleisch@augen-franziskus.de

[1] Department of Ophthalmology, St. Franziskus-Hospital, Hohenzollernring 74, 48145 Muenster, Germany

[2] M3 Macula Monitor Muenster GmbH & Co KG, Muenster, Germany

[3] Westphalia DataLab GmbH, Muenster, Germany

[4] Department of Ophthalmology, University Duisburg-Essen, Essen, Germany

[5] Achim Wessing Institute of Ophthalmic Diagnostic, University Duisburg-Essen, Essen, Germany

[6] NIHR Biomedical Resource Centre, UCL Institute of Ophthalmology and Moorfields Eye Hospital NHS Trust, London, UK of Great Britain and Northern Ireland

Key messages

- Decisions for treatment of neovascular AMD are often challenging for non-retina specialists.
- Initial and repeated indication of anti-VEGF therapy in neovascular AMD can be assisted using deep learning network analysis.
- The algorithm can be supervised by activation map volume scan visualization.

Abbreviations

AI	Artificial intelligence
AMD	Age-related macular degeneration
AUC	Area under curve
BCVA	Best corrected visual acuity
BM	Bruch's membrane
CATT trial	Comparison of Age-related Macular Degeneration Treatment Trials: Lucentis-Avastin Trial
CNN	Convolutional neural network
CNV	Choroidal neovascularization
FA	Fluorescein angiography
ILM	Inner limiting membrane
IVAN trial	Inhibition of VEGF in Age-related choroidal Neovascularisation trial
GPU	Graphics processing unit
LSTM	Long short-term memory
M	Mean score
nAMD	Neovascular age-related macular degeneration
PRN	Pro re nata
RC	reading centre
RPE	Retinal pigment epithelium
ROC	Receiver operating characteristic
ReLU	Rectified linear unit
SD	Standard deviation
SD-OCT	Spectral domain optical coherence tomography
tanh	Hyperbolic tangent
TNR	True negative rate
TPR	True positive rate
VEGF	Vascular endothelial growth factor

Introduction

Anti-VEGF therapy is currently the standard for the treatment of neovascular age related macular degeneration (nAMD) [1]. In all prospective studies the minimal inclusion criteria was "occult (type 1) choroidal neovascularization (CNV) with recent disease progression". But analyzing the individual clinical nAMD requiring anti-VEGF therapy using fundus examination, fluorescein angiography (FA) and spectral domain optical coherence tomography (SD-OCT) in real-life, a misdiagnosis and disagreement between treating doctors and reading centres in a range between 5 and 18% could be identified [2, 3]. Therefore, it is a clinical need to improve the decision process for anti-VEGF treatment and retreatment of nAMD.

Recent years have seen a rapid implementation of artificial intelligence (AI) in medical image analytics and potential treatment predictions [4–9]. They have been established in subcortical vascular cognitive impairment [10] and glaucoma [11]. Also, in medical retina these AI approaches have been shown reliable to differentiate between different macular diseases [7, 8, 12–14]. In addition, previous AI studies in nAMD have shown an acceptable prediction for conversion of nAMD in the same eye [9, 15] and the second eye [16]. Also the differentiation of OCT images between normal vs. pathological findings (AMD) [17] as well as the characterization of specific OCT biomarkers [15–18] could be achieved using AI algorithms. In this study, we aimed to develop an AI-based decision support for non-retina specialists in daily clinical work (see Fig. 1). Two experiments were carried out for this purpose. The first experiment aims to differentiate between nAMD patients who need anti-VEGF therapy from those AMD patients who do not. The second experiment works on facilitating retreatment decisions (stabilised vs. active nAMD decision) during follow up. In both situations, referral to a treatment centre would be recommended. To demonstrate the robustness, the algorithms were tested via cross-validation and benchmarked against multiple retina specialists.

The applicability of the approach is underlined by the fact that no specific OCT features were extracted or annotated, that an end-to-end process was established, that the trained models were based on image data taken from daily routine treatment, and that special requirements for the images, such as scan density, were left out and thus the developed AI model can be more easily used for clinical application.

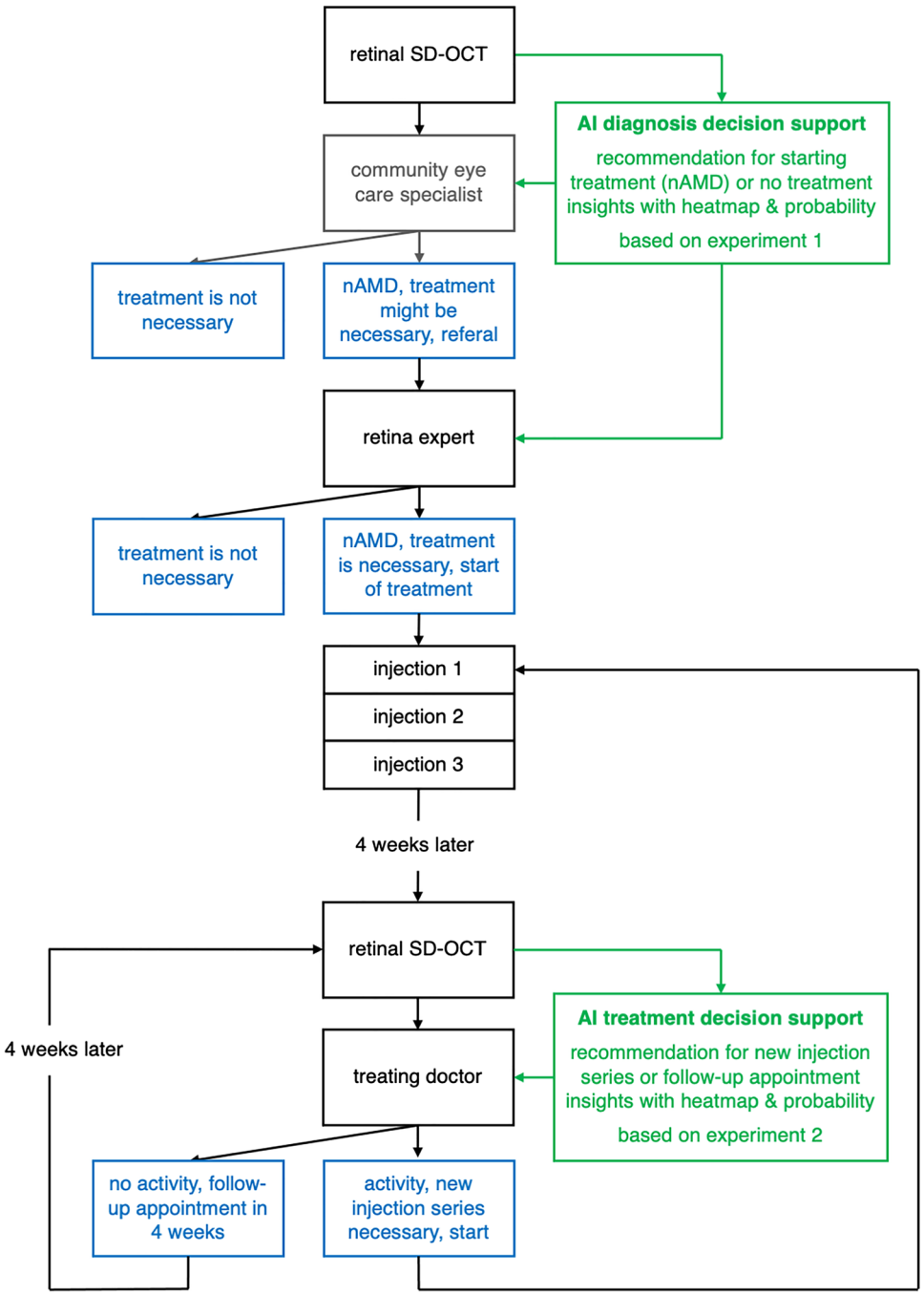

Fig. 1 Treatment procedure for nAMD following the PRN schema with AI decision support systems for initial indication and retreatment decision

Patients and methods

Overview

OCT scans and treatment decisions were collected during daily practice. For a retrospective cohort of patients without previous selection where at least one eye was treated following a standardized treatment protocol, this data was used as input data. For experiment 1, the two classes are fellow eyes without indication for treatment and eyes requiring treatment. For experiment 2, the two classes consist of the doctor's assessments of stabilised nAMD or active nAMD during the course of treatment. Only SD-OCT scans with a standardised resolution made by Heidelberg Engineering devices were used.

A single data preprocessing pipeline and for each experiment a convolutional neural network (CNN) applying deep learning were developed. Preprocessing consisted of normalizing image eye side orientation, downsampling to a quarter of the original resolution, removing areas outside a

defined region of interest (ROI) and contrast enhancement. To increase the amount of training data, the dataset was augmented by variations of the original images randomly rotated and shifted. 3D convolutional blocks were used in the CNNs so that the models are trained by all dimensions of the OCT volume. Experiment 1 uses one OCT scan and its target value in a single CNN. In experiment 2, two subsequent OCT scans of one eye and the corresponding decision for the latter image were used. Both inputs were processed by one CNN and their separate outputs combined using a LSTM to also capture temporal information.

To demonstrate the robustness of the developed algorithms, cross-validation (tenfold) was used. In addition, we generated saliency maps of the deep learning model to visualize the relevant characteristics of the individual deep learning analysis and results of the algorithms. These saliency map characteristics of initial indication decisions were analysed by retina specialists (H. F., B. H.-B., M. Z.) for corresponding biomarkers.

To benchmark the AI analysis, the results were compared to gradings made by retina specialists (B. H.-B., M. Z., M. G.) for differentiation of initial indication of patient eyes.

Data

The Department of Ophthalmology, St. Franziskus-Hospital, Muenster, Germany, has established a digital platform between local ophthalmologists and its clinical treatment centre for cooperative anti-VEGF treatment of patients with nAMD. Using this platform, all images and clinically relevant information are exchanged digitally prior to initial treatment and before every subsequent treatment [19] with intravitreal anti-VEGF therapy. Decisions for treatment and retreatment were based on reading centre (RC) analysis at the treatment centre (RC: M^3 Macula Monitor Muenster GmbH & Co KG, Muenster, Germany). The study used the pro re nata (PRN) Inhibition of VEGF in Age-related choroidal Neovascularisation (IVAN) [20] trial protocol (three monthly injections). Treatment and retreatment decision were defined following the internationally published criteria (Comparison of Age-related Macular Degeneration Treatments Trials: Lucentis-Avastin Trial [21], IVAN trial [20]).

Using this cooperative analysis and treatment system, a consecutive unfiltered cohort of 1503 nAMD patients with SD-OCT volume scans and clinical information was analysed. Patients were seen between 2012–2020. Clinical information (best corrected visual acuity (BCVA), FA, gender) and SD-OCT volume scans (Spectralis SD-OCT 1 or 2, Heidelberg Engineering, Heidelberg, Germany, 49 B-scans, $20° \times 20°$) were collected. SD-OCT images of fellow eyes were also transferred to the RC and were used as a comparative cohort. These eyes demonstrated most often early/intermediate AMD, but a substantial number of eyes also had disciform scars with BCVA > 1.3 logMAR or additional other pathologies like epiretinal gliosis (Table 1). The study was conducted in compliance with the Declaration of Helsinki. Ethics Committee (University of Muenster) approval was obtained.

Artificial intelligence is based on experience encoded in data. To develop the AI decision support algorithms, we generated two data sets from this cohort that contain the historical imaging data from SD-OCT volume scans of AMD-affected patients and their corresponding treatment decisions from retina specialists. We used these data sets to train and test the algorithms.

The historical SD-OCT image data and meta data were extracted from Heidelberg Engineering's HEYEX 2 software, which uses a proprietary data format. These files contain the raw pixel data of the SD-OCT scans, in our case with 49 B-Scans containing 512 A-Scans with 496 pixels. Additionally, the file's meta data contain SD-OCT segmentation lines automatically generated by the HEYEX 2 software. The historical patient treatment data at every examination date was extracted from a structured medical record system. The predefined treatment process supported by the medical

Table 1 Breakdown of "no treatment" class into subclasses for experiment 1 by expert opinion

Definition of subclass	Number of predictions		Class prevalence (sum)	Subclass TNR (specificity)
	Initial treatment	no treatment		
Early AMD	3	137	140	98%
Intermediate AMD	32	211	243	87%
Geographic atrophy	12	52	64	81%
Disciform scar	18	34	52	65%
Other pathologies (e.g. epiretinal membrane, pattern dystrophy)	8	26	34	76%
Namd with bcva > 1.3 logmar	62	76	138	55%
Not graded (missing or low-quality data)	3	3	6	50%
Totals	138	539	677	80%

TNR true negative rate

record systems ensures treatment process integrity and the use of structured treatment forms ensures high data quality.

We linked the image and treatment data for each patient based on the image acquisition date and the medical record date.

Data set for experiment 1: no treatment vs. initial treatment

To develop an AI decision support algorithm that differentiates between no treatment vs. initial treatment of suspicious nAMD cases, we selected all SD-OCT volumes of initial RC examinations with a nAMD indication and a succeeding intravitreal anti-VEGF therapy resulting into 1712 eyes with nAMD that required anti-VEGF treatment. SD-OCT images of fellow eyes without an indication for anti-VEGF-therapy were used as a comparative cohort to form the no treatment class. The no treatment class contained 737 eyes. All samples of this class were evaluated by retina specialists to divide it into six subclasses for different stages of AMD and other pathologies (early AMD, intermediate AMD, geographic atrophy, disciform scars, nAMD with BCVA > 1.3 logMAR, other pathologies).

We ensured that only the very first indication of one patient's eye was included in our data set since there were patients with multiple AMD indications with treatment gaps of several years. Overall, this unfiltered data contained 2449 eyes from 1503 patients.

Finally, after filtering for sufficient segmentation lines, 2322 eyes of 1477 patients (1644 eyes with nAMD that required anti-VEGF treatment and 678 eyes where no treatment was indicated) were considered in the following experiment. This data underwent the preprocessing steps and was used for training.

Data set for experiment 2: stabilised nAMD vs. active nAMD

The treatment following the IVAN trial protocol makes ongoing AMD examinations of activation criteria inevitable. Ophthalmologists decide about retreatment with a new anti-VEGF injection series. To develop a decision support algorithm that helps differentiating between stabilised vs. active nAMD, we assembled a data set that contains historical SD-OCT volumes and the corresponding retreatment decision. When following the PRN treatment schema, the decision can either be retreatment (active nAMD class) resulting into a new anti-VEGF injection series or follow-up visit resulting in a new examination four weeks later (stabilised nAMD class). We selected every two consecutive SD-OCT volume scans of one initially treated unique patient eye's treatment history and the corresponding retreatment decision.

For example, from the following ordered images for one patient eye $SD\text{-}OCT_{t\text{-}3}$, $SD\text{-}OCT_{t\text{-}2}$, $SD\text{-}OCT_{t\text{-}1}$, $SD\text{-}OCT_{t0}$ three unique timeseries-samples were generated:

Timeseries sample 1: $SD\text{-}OCT_{t\text{-}3}$, $SD\text{-}OCT_{t\text{-}2}$, retreatment decision t_{-2}
Timeseries sample 2: $SD\text{-}OCT_{t\text{-}2}$, $SD\text{-}OCT_{t\text{-}1}$, retreatment decision t_{-1}
Timeseries sample 3: $SD\text{-}OCT_{t\text{-}1}$, $SD\text{-}OCT_{t0}$, retreatment decision t_0

By providing two consecutive SD-OCTs to the CNN, the network can learn to compare both volumes to make a decision.

We also run experiments with only one SD-OCT volume but found out that the AI performance increases by learning from two consecutive SD-OCTs as seen in the "Results" section. This coincides with how retina specialists evaluate the development of activation criteria by examining the preceding and current SD-OCT scans.

In total 9451 timeseries samples containing two consecutive SD-OCT volumes were built: 5717 SD-OCT volume scan pairs with decision of stabilised CNV were compared with 3734 SD-OCT volume scan pairs with decisions for retreatment. Only patient eyes and their follow-up appointments which previously had been given an initial diagnosis of nAMD needing treatment (see data set 1) appeared in this dataset.

Data preprocessing

To aid model training, we evaluated several, appropriate image preprocessing methods and chose the most effective for both experiments. The contribution of each preprocessing step to the model performance for experiment 1 can be found in the result section and in Table 2. Figure 2 shows the steps of the final data preprocessing pipeline with one sample slice. Raw data of pixel-wise reflectivity of the SD-OCT scans were separated and manually transferred into the data preprocessing pipeline. For the analysis, SD-OCT scans with 49 B-Scans containing 512 A-Scans with 496 pixels were filtered from the obtained dataset (volumes with the

Table 2 AUC results for experiment 1 and different preprocessing steps

Preprocessing	CV AUC
downsampled	0.880
downsampled, ROI	0.906
downsampled, ROI, CLAHE	0.925
downsampled, ROI, CLAHE, augmentation (final)	0.927
fullsize, ROI, CLAHE	0.894

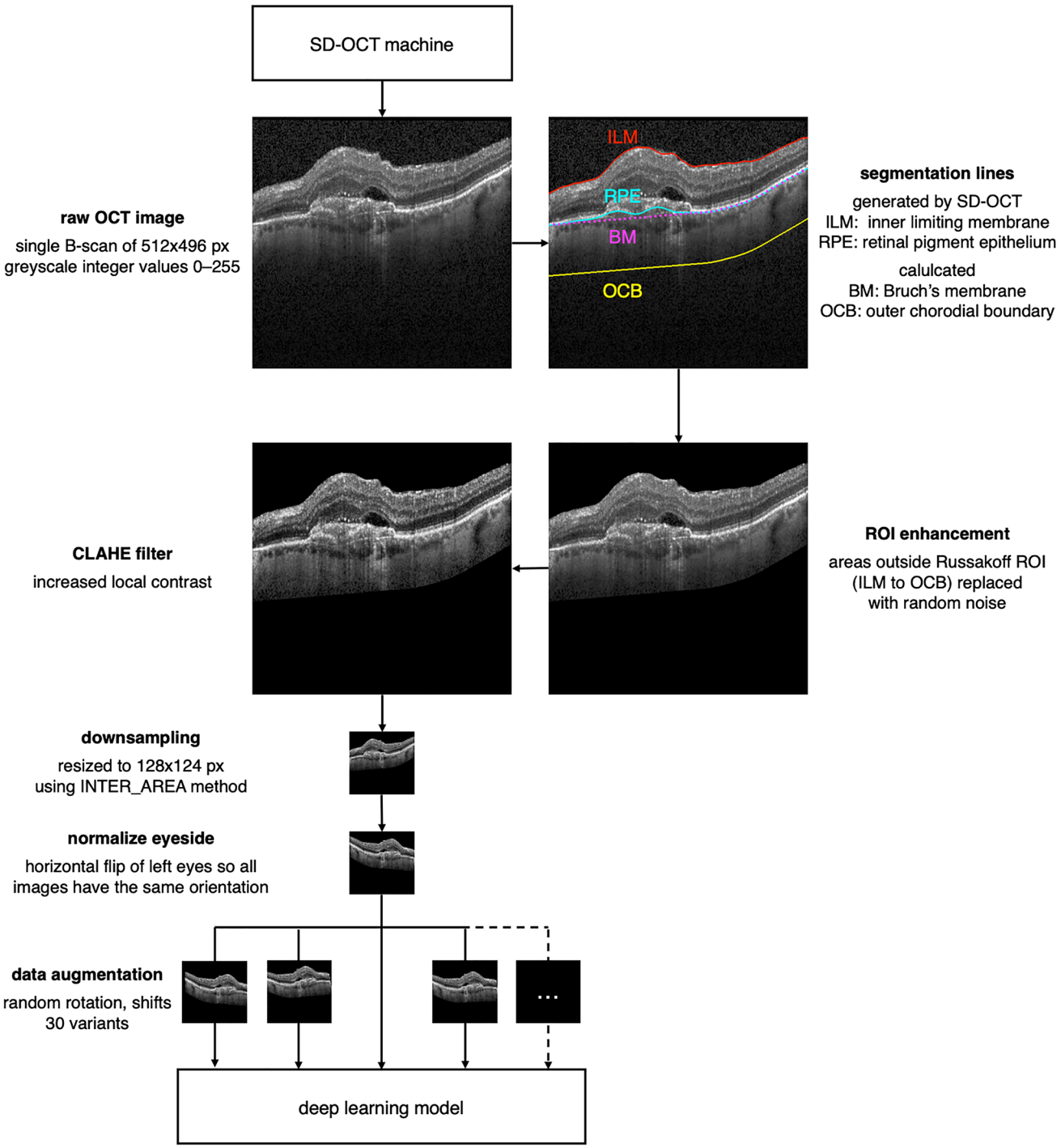

Fig. 2 Final preprocessing pipeline with one sample slice from the raw image data extracted from the SD-OCT machine to the final input used to train the deep learning model

dimensions 512 × 496x49). Before feeding the images to the deep learning model, the provided images underwent preprocessing. No SD-OCT scans were excluded due to image quality, only a small fraction (up to 7 percent) with nonexistent or highly discontinuous segmentation lines was disregarded, as they were used for the next preprocessing step.

A region of interest (ROI) that is considered prognostic of AMD like in Russakoff et al. [9] was defined so that the CNN focuses on relevant areas only and variance in the dataset is reduced. For this, the area between the ILM segmentation line and the lower bound of the choroid area (outer choroidal boundary, OCB) is automatically identified. The areas outside

of this ROI (the vitreous body above the inner limiting membrane (ILM) and the sclera below the choroid) were replaced with 97% black and 3% low intensity (grey values of 1–64) random noise pixels to improve saliency map interpretability. The ILM segmentation line is produced by the SD-OCT proprietary software. We used the provided retinal pigment epithelium (RPE) segmentation line and generated the convex hull around RPE as an estimation of the Bruch's Membrane (BM) [22]. Following to Russakoff et al. [9] we shifted this BM line down in parallel by 390 µm (empirical mean + 2 SD of the subfoveal choroidal thickness in a population with AMD) to define a lower bound of the ROI.

In the next step, contrast enhancement was applied to the images by using contrast-limited adaptive histogram equalization (CLAHE).

Finally, the dimensions of the B-scans were downsampled to 128 × × 124 by using OpenCV's interpolation method INTER_AREA [23], resulting into a volume of 128 × 124 × 49. As the scaling factor 4 is a common divisor of the original dimensions, each resized pixel intensity shows the average of 4 × 4 pixels in the original image. Image downsampling is a common feature in deep learning for ophthalmic image analysis [9, 24]. Downsampling has been done in both aspect ratio conserving [24] and non-conserving for both OCT [9] and for fundus image analysis [24]. Lower resolution images as model input allow for faster model training and parameter tuning in development and use less hardware resources both in training as well as inference. In addition, it increases transferability of the model to inputs by other SD-OCT machines with varying resolutions, vendor-specific differences in texture granularity and visual artefacts. To verify this downsampling does not significantly affect model performance, we conducted a ceteris paribus comparison for experiment 1 with an adapted CNN design to account for the bigger input dimensions.

To have a more uniform dataset, all images were normalized regarding their horizontal orientation relative to the nose, meaning images from left eyes were flipped to have the same orientation as right eyes. To generally enlarge the training data, compensate for natural variations in scan positioning and alleviate overfitting, the training data was augmented by random rotation (5–10°), vertical shift (3–15%), and horizontal shift (3–10%). We rescaled all pixel values of 0–255 to floats of 0.0–1.0 to improve the model training convergence speed.

All models were trained end-to-end, without any prior segmentation or biomarker definition.

Deep learning

Both algorithms were trained using end-to-end deep learning, without any prior segmentation or biomarker definition. Two new deep learning architectures were developed.

Architecture experiment 1: no treatment vs. initial treatment

The 3D CNN scheme for experiment 1 consists of three stacked convolutional blocks followed by a global average pooling and a fully connected dense layers with rectified linear unit (ReLU) as the activation function. Finally, a softmax layer yields class probabilities for the input volume. Each convolutional block is composed (of a sequence) of a 3D convolutional layer, ReLU activation, batch-normalization and a 3D max pooling layer. Table 3 summarizes the structure and hyper-parameters of the network.

To mitigate overfitting, we applied L2-regularization (lambda = 0.005) in the convolutional layers and dropout in the fully connected layer with a dropout rate of 0.5. Furthermore, early stopping policy terminated the training once the monitored validation loss had not improved for multiple epochs. For the final model the weights of the epoch with best performance (lowest validation loss) were selected.

Architecture experiment 2: stabilised nAMD vs. active nAMD

In experiment 2 each sample is treated as a timeseries of two SD-OCT scans, containing the current and the previous scan from a single patient and eye. Since the input contains spatial and temporal information, a hybrid model involving a CNN and long-short term memory (LSTM) was implemented. LSTM is a proven class of model in deep learning used to process sequence of data. In the proposed model CNN is applied to extract the feature vector representation from each of the SD-OCT scans, passing the resulting feature vectors to the LSTM for the sequence

Table 3 Parameters of the 3D-CNN architecture in experiment 1

Layer	Units	Kernel Size	Activation	L2
3D convolution_1	32	3 × 3 × 3	ReLU	0.005
Batch normalization_1				
3D Max pooling_1		2 × 2 × 2		
3D convolution_2	32	3 × 3 × 3	ReLU	0.005
Batch normalization_2				
3D Max pooling_2		2 × 2 × 2		
3D convolution_3	32	3 × 3 × 3	ReLU	0.005
Batch normalization_3				
3D Max pooling_3		4 × 4 × 4		
Global Average Pooling				
Fully Connected	64			
Dropout (30%)				
Fully Connected	2		Softmax	

(L2) = L2-regularization

(ReLU) = rectified linear unit

learning of the above mentioned timeseries. This model architecture was comprised by the 3D CNN architecture from experiment 1 (here with lambda = 0.0001) as a time-distributed input to an LSTM layer with 64 hidden cells outputting only the last hidden cell with activated internal dropout-rate and a recurrent-dropout-rate both of with 0.1, and hyperbolic tangent (tanh) as the activation function. The output of the LSTM layer is connected to a fully connected layer with 64 units, and a dropout layer with a dropout-rate of 0.3 concluding to a final softmax layer for the two-class prediction problem.

Training

For training, the whole dataset was first randomly shuffled. To get a reliable evaluation of the model performance, we conducted tenfold cross validation at patient level. In each of the 10 training iterations a new rotating subset with 10% of all samples was held out for the test set. This ensured that each sample was classified once as part of a test set. The remaining samples were randomly divided into training (72% of all samples) and validation set (18%). To address data leakage in each iteration all data relating to a patient appeared strictly in one subset only. The validation sets served for early stopping and best model selection in each iteration. For overall AUC of an experiment, the mean value of the AUCs from all 10 tests sets was calculated.

Both models were trained by Nadam optimizer [25], with an initial learning rate of 0.001 using cross entropy as the loss function. In experiment 1 the initial learning rate of 0.001 was adapted during training to 0.0001 after the 10^{th} epochs and then to 0.00001 after 20^{th} epoch. Similarly, in experiment 2, after 20^{th} epoch we set the learning rate to 0.005 and to 0.0025 after 30^{th} epoch. The batch size was set to 4. We assessed the prediction performance based on the area under receiver operating characteristic curve (AUC) score. An AUC of 1 indicates a perfect classifier, while 0.5 represents a classifier without discriminative power. The receiver operating characteristic curve (ROC) itself plots the relation between the true positive and false positive rate. In this study, we preferred using 3D CNN over 2D CNN topologies, to also capture the spatial context in the B-scans dimension.

A special platform was created for configuring and validating the model parameters, tracking the experiments, visualizing the results and evaluating the performance. Keras [26] served as the deep learning framework using TensorFlow [27] as the backend. The experiments ran on a dedicated machine running Ubuntu Server 20.04 and equipped with two linked GPUs (Nvidia GeForce Titan RTX, NVIDIA Corporation, Santa Clara, USA).

Saliency map viewer

In addition, a saliency map viewer was developed to visualize the relevant characteristics of the individual deep learning analysis and results of the algorithms using colour coding. Saliency maps are obtained by computing the partial derivatives of the output class score with respect to each input image pixel. The magnitude of these partial derivatives denotes the contribution of each pixel to the predicted class [28, 29]. For improved interpretation a gaussian filter with a standard deviation value of 0.8 is applied to smooth out the resulting/calculated pixel values of the saliency map. Highly activated areas are highlighted in red to yellow colour.

Grading by retinal specialists

To compare our results with human decision making, we let three retina specialists perform a grading of 102 randomly chosen samples. Each grader was given the original full resolution SD-OCT volume scan used in the initial indication without any additional clinical information to differentiate between treatment and no treatment.

Results

Experiment 1: no treatment vs. initial treatment

In experiment 1, besides the final scores, we also determined the effects of the different preprocessing steps to evaluate their usefulness for the model. Without any preprocessing except resizing each B-scan to 1/4th of the original resolution a model was trained with an AUC of 0.880 to serve as a baseline for iteratively evaluating the usefulness of further preprocessing steps. All values were recorded with tenfold cross validation. By utilizing the ROI enhancement preprocessing step after resizing the AUC increased to 0.906. Additionally, applying CLAHE, the mean AUC improved to 0.925. To verify that our downsampling did not significantly affect model performance, we conducted a ceteris paribus comparison for the preprocessing pipeline with ROI enhancement and CLAHE applied but using full-sized images with an adapted CNN design to account for the bigger input dimensions. This showed that using full-sized images and the resulting bigger variance in samples produced lower AUC of 0.903 (SD: 0.018), indicating that our sample size did not suffice for the increased number of features in the full-size image. By extending the preprocessing pipeline of ROI enhancement, CLAHE and downsampling with augmentation, the final AUC showed a slight improvement: The model for initial indication achieved a mean AUC of 0.927 (standard deviation (SD): 0.018). Figure 3 depicts the single ROCs, the mean ROC and the standard deviation

of all ten runs. Additionally, an operating point for the optimal operating threshold according to Zweig and Campbell [30] with equal costs for all decisions (m = 1, so TPR-TNR is maximized) is given. Also, the frequency of the prediction value was analysed to evaluate the effectiveness of the network (Fig. 4). Among all instances the model predicts with high confidence the correct class with only small portion of misclassifications. Especially for true predictions of initial treatment a high frequency of confidence values close to 1.0 was observed, while most true predictions of no treatment had a confidence value of at least 0.8. This validates the model's capacity discriminating no treatment versus initial treatment decisions with high confidences.

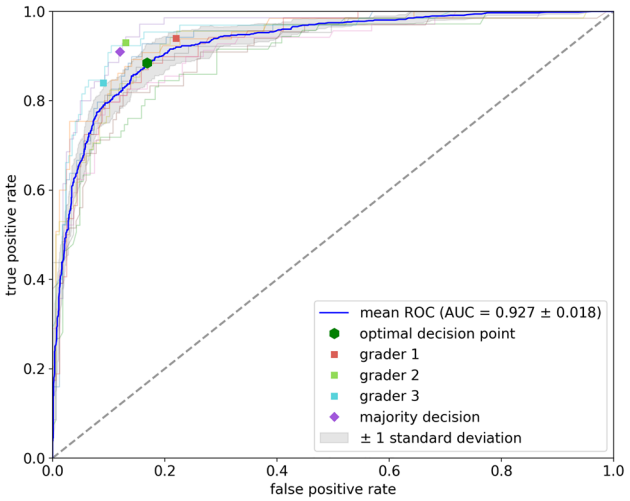

Fig. 3 Illustrates the receiver operating characteristic curves (ROC) for experiment 1, the faint-coloured lines show each of the 10 folds, the thick blue line the mean of all experiments; area under receiver operating characteristic curve (AUC)

To further understand the model performance, all samples of the no treatment class were grouped by their respective subclass as described in the Data section. The number of correct "no treatment" (true negative) and incorrect "initial treatment" (false positive) predictions for the default decision threshold of 0.5 as well as the true negative rate (TNR or specificity) per subclass can be seen in Table 1. For samples with no treatment AMD the model showed the highest subclass TNR of 91% across both classes. Especially eyes with BCVA > 1.3 logMAR, where treatment is generally not considered, leads to a low subclass TNR of 55%. Even with this irregular real-life dataset, a big majority of patients requiring no treatment were correctly predicted as such, with a true negative rate (TNR) of 80%. When pruning the no treatment class by removing all subclasses except early/intermediate AMD, model performance could be improved

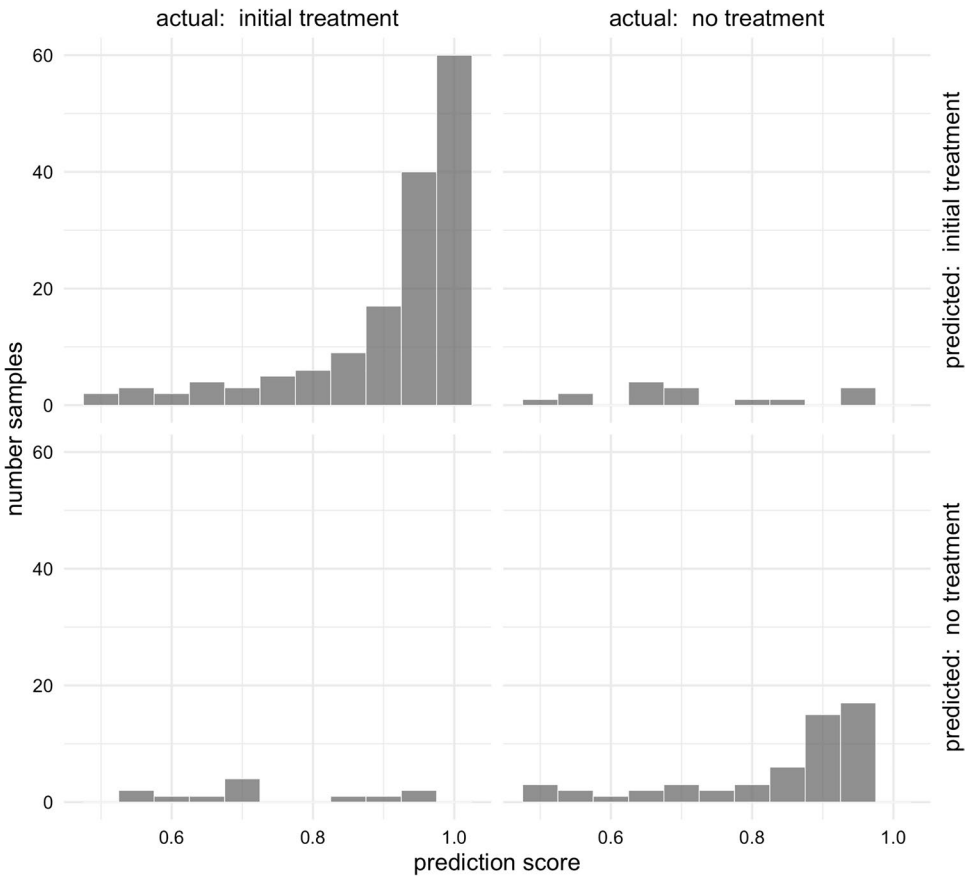

Fig. 4 Frequency of the prediction value no treatment and initial treatment of AMD

significantly to a TNR of 97%. The mean AUC increased from 0.927 with real-life data to 0.976 with pruned data. This indicates that improvements for real-life applications could be reached by automatic filtering of known properties (like BCVA) or using a multiclass model which differentiates between characteristic subclasses.

Experiment 2: stabilised nAMD vs. active nAMD

Using the dataset without augmentation but with the final preprocessing pipeline the model for differentiation of stabilised vs. active nAMD achieved a mean AUC of 0.842 (SD: 0.022). By applying augmentation, the performance increased to a mean AUC of 0.865 (SD: 0.027; Fig. 5), which is the final AUC for experiment 2.

We were also interested to assess the benefit of utilizing preceding and current SD-OCT as a timeseries against the case of only using the current SD-OCT as input. For the case of using a single SD-OCT volume as input, the deep learning model consisted of the 3D-CNN part of our 3D-CNN-LSTM architecture only. For this comparison, datasets without augmentation were used. The model with the single (current) SD-OCT volume achieved an AUC of only 0.815 (SD: 0.027), compared to the AUC of 0.842 (SD: 0.022) in the timeseries case using the LSTM architecture.

Also, the frequency of the prediction value was analysed to evaluate the effectiveness of the network (Fig. 6). For true predictions of stabilised nAMD a high frequency of confidence values close to 1.0 was observed, while most true predictions of active nAMD had a confidence value of at least 0.8. This validates the model's capacity discriminating stabilised nAMD versus active nAMD decisions with high confidences.

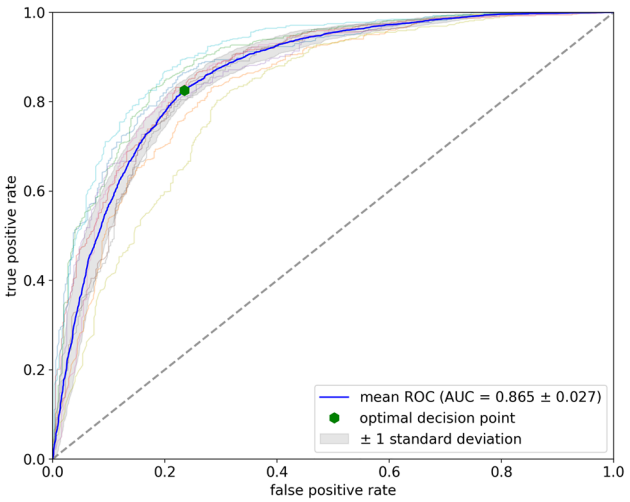

Fig. 5 Illustrates the receiver operating characteristic curves (ROC) for experiment 2, the faint-coloured lines show each of the 10 folds, the thick blue line the mean of all experiments; area under receiver operating characteristic curve (AUC)

Saliency map analysis

Figure 7a shows the saliency map for a single B-scan direction with highly activated areas in red to yellow colour. Figure 7b is showing the saliency map in direction across all 49 B-scans. Since the areas, which demonstrated activation, are continuous between adjacent B-scans, it is indicating the value of using 3D CNN instead of 2D CNN. In the 3D CNN different structures (interface vitreous/retina, subretinal, intraretinal, sub-RPE space and choroid) could be differentiated. To define a gradation of the relevant structures, on which the algorithm decided towards an individual recommendation (red coded structure), the saliency maps of 24 patients with a predicted probability of ≥ 97.5% and an active stage of the nAMD were analysed. Scores from 0 to 2 (0 = no staining, 1 = slight staining, 2 = intensive staining) were used for each morphological structure and a mean score (M) was registered. This analysis of colour intensity on individual saliency maps was applied on complete volume scans by two independent graders using standard images for classification. The retina/vitreous interface was the most important structure relevant for the activity decision of the algorithm (M = 2.0; SD ± 0). This is followed by the subretinal space (M = 1.375; SD ± 0.770), the intraretinal cysts (M = 1.0; SD ± 0.933), the sub-RPE space (M = 0.667; SD = 0.868) and the choroid (M = 0.625; SD ± 0.824). Therefore, using the saliency map analysis, the deep learning model could visualize areas in the SD-OCT images, which are relevant for an individual decision and therefore the results of the AI algorithm can be correlated with typical corresponding retinal AMD changes.

Comparison with retinal specialists

The metrics of manual grading can be seen in Table 4. The results for grading with only SD-OCT volume information available show a high interrater reliability with a Fleiss' Kappa [31] of κ = 0.824. As ground truth, the decisions by doctors in our real-life dataset were used and compared to the majority vote of the three retinal specialists. A Cohen's Kappa value of κ = 0.776 was observed. Sensitivity for each grader ranged from 78 to 94% (majority vote: 91%), specificity from 78 to 91% (majority vote: 87%). All false evaluations by majority vote were looked at manually: the 6 false negatives can be explained by unicus situation and activity which is only visible in other imaging modalities than SD-OCT images, while the remaining 4 false positives either had BCVA > 1.3 logMAR or disciform scars.

The grading performance can be compared to our predictions made in tenfold cross-validation for these 102

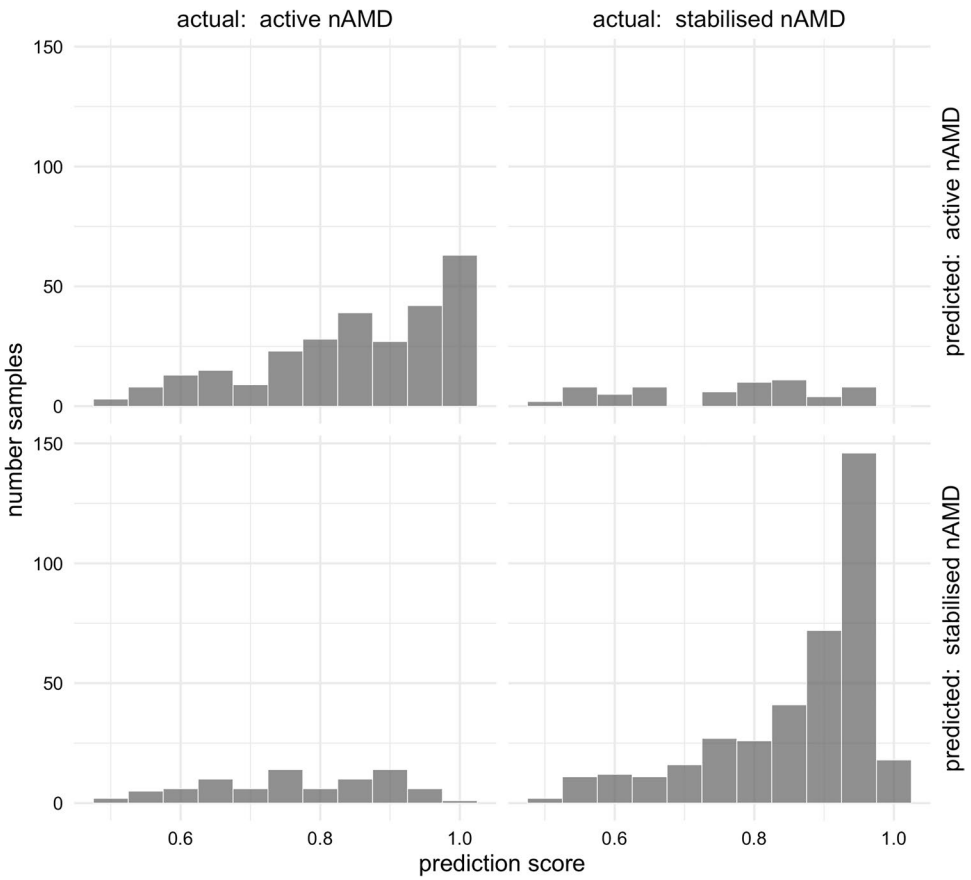

Fig. 6 Frequency of the prediction value stabilised nAMD and active nAMD

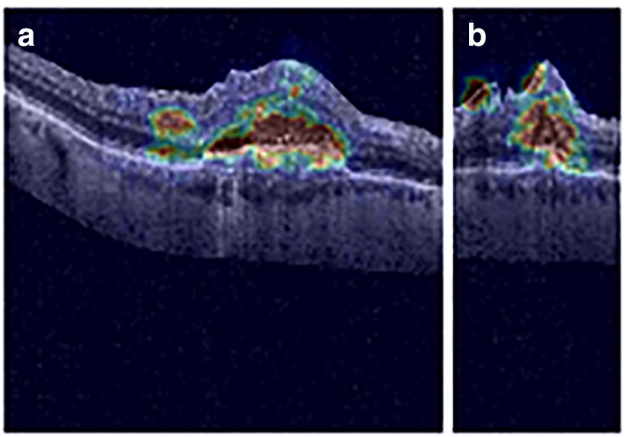

Fig. 7 Saliency map of one sample OCT for a single B-scan (**a**) and in z-axis direction across 49 B-scans (**b**)

samples as they are based on the same image data. Again, the doctors' clinical decisions were used for comparison. With the default decision threshold of 0.5 a Cohen's Kappa of $\kappa = 0.650$ was observed for all model predictions, being close to human performance.

Discussion

In this study using an unspecified real-life cohort of nAMD patients two new CNNs have been developed, which can support non-retina specialists to distinguish between AMD cases with no treatment needed and treatment indicated nAMD as well as between stabilised and retreatment indicated situations based on SD-OCT raw data. These algorithms can be applied to daily practice to support the decision of non-retina specialists for referral to treatment centres. The defining characteristics of these algorithms are end-to-end processing and their independence of specific OCT feature analysis. In addition, the saliency map viewer could visualize the relevant characteristics for the algorithms. In previous studies the developed AI algorithms were predominantly addressing the question of AI-assisted automatically segmentations on SD-OCT images [32, 33]. In additional AI studies on nAMD the prediction for conversion from intermediate into nAMD was of major interest [9, 14, 15], but also the analysis for predictive biomarkers for AMD progression from intermediate AMD into nAMD was in the focus of interest [18, 34]. Especially the AI analysis of fluid distribution during anti-VEGF therapy of nAMD could be successfully achieved [35]. This study focuses on developing AI algorithms differentiating between no treatment

Table 4 Metrics of clinical experts in grading

Expert decision based only on SD-OCT								
Needing initial treatment?	Grader 1		Grader 2		Grader 3		Majority	
	Yes	No	Yes	No	Yes	No	Yes	No
Yes	66	4	65	5	59	11	64	6
No	7	25	4	28	3	29	4	28
Cohen's Kappa	0.743		0.797		0.702		0.776	

and treatment (initially and retreatment) in intravitreal anti-VEGF therapy in nAMD.

In these and other AMD studies [15, 16, 36] an AUC of > 0.80 was considered as a clinically good and meaningful differentiation. The results of the present study with an AUC of 0.927 for the differentiation between treatment-indicated nAMD and fellow eyes with AMD cases with no treatment needed can therefore be considered clinically relevant, especially because the control group of fellow eyes contained beside eyes with early and intermediate AMD, a considerable number of eyes with late stage nAMD and other pathologies. Also, the AUC of 0.865 for the differentiation between stabilised and retreatment-indicated nAMD are in this relevant range. The clinical relevance of these results is also highlighted by the fact, that in both situations in the IVAN and CATT trial there was also a disagreement between treating retina specialist and the RC of approximately 20% [20, 21]. Because the developed AI algorithms were based on unselected real-life treatment data and because they demonstrated robustness against downsampling, cross-validation, and retinal specialist´s opinion, these algorithms appear to be valid to be tested as a decision aid for referral in clinical practice.

In addition, the developed saliency map viewer could visualize the relevant characteristics of an individual deep learning analysis using colour coding the prediction of the trained 3D CNN models. In initial indication decisions with a predicted probability of ≥ 97.5% and an active stage of the nAMD, the retina/vitreous interface was the most important structure relevant for the activity decision of the algorithm, which may be a characteristic for retina thickness. Furthermore, changes in the subretinal space representing subretinal fluid, intraretinal cysts, sub-RPE fluid and in some SD-OCT scans analysis changes in the choroid were relevant. Therefore, using the saliency map analysis, the deep learning model could visualize areas in the SD-OCT volume scan, which supports the AI decision aid by visualizing the basic structural correlate for the examining ophthalmologist.

Our downsampling of each of the 49 B-scans of one SD-OCT volume to 1/4th of the original dimensions might have led to information loss in the related biomarkers, yielding in decreased model performance. Comparison experiments using the developed model architecture showed that the full-sized volumes decreased scores against expectation. However, the model was not fully optimized for full-size volumes and the sample size might be too small for the increased number of features. In everyday clinical practice retina specialists base their diagnostic decision also on additional information, such as fundus images, BCVA, patients age and activity criteria which could be integrated into a clinical decision aid. Judging the treatment using an algorithm based on OCT alone, may be more suitable for patients with large lesions and excessive exudation.

The cohorts used in this study were data of unselected case series of the clinical routine in the Department of Ophthalmology, St. Franziskus-Hospital, Muenster. Therefore, for retreatment some individual SD-OCT images were considered as stabilised in which the treatment was terminated because further anti-VEGF treatment was not considered to improve the situation. Eliminating these cases and re-evaluating all decisions from the learning cohort as well as increasing in the number of SD-OCT volume scans by developing automatisation method for SD-OCT-data transfer may result in significant further qualitative improvement of individual predictions. Even though the saliency map focused clinically relevant areas, they should be interpreted with caution, since data set was small in relation to the diversity of patterns in the images.

In summary, the results of our study demonstrate, that the developed AI algorithms can have great implications for the future development of medical care models between non-retina and retina specialists in the treatment of patients with nAMD in real-life clinical practice. These models also offer the possibility of being extended to collaborations between non-physician providers and retina specialists. The analysis of SD-OCT scans of AMD patients with initial or repeated indications for anti-VEGF therapy in nAMD using this algorithm may support non-retina specialists in their decision for referral to a treatment centre. In addition, the individual decision of the algorithm can be supervised by saliency map volume scan visualization. This algorithm can therefore improve the performance and accuracy of non-retina specialists in real life to achieve reading centre standard.

Acknowledgements H. Faatz, MD [1,6], as a retina specialist, he was involved in analysing the saliency maps. B. Heimes-Bussmann, MD [1,6,

M. Ziegler, MD [1,6], as retina specialists, they were involved in benchmarking the AI analysis of the initial indication and in analysing the saliency maps.

Funding This study was funded by Novartis Pharma GmbH, Nuernberg, Germany. The sponsor or funding organization had no role in the design or conduct of this research.

Declarations

Ethics approval All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Approval for the study was obtained from the local ethics committee at the University of Muenster.

Informed consent For this type of study, formal consent is not required.

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript. O. Ester, S. Aydin, M. Quassowski and K. Rothaus declare they have no financial interests. M. Gutfleisch received speaker honoraria from Novartis, Bayer and Zeiss outside the submitted work. G. Spital received speaker honoraria from Zeiss, OD-OS and Allergan outside the submitted work. A. Lommatzsch received speaker honoraria from Novartis, Bayer and Zeiss outside the submitted work. A. M. Dubis has a patent OCT analysis technology pending, and a patent prediction method from retinal imaging pending outside the submitted work. D. Pauleikhoff received speaker honoraria from Bayer, Novartis, Zeiss and Allergan outside the submitted work.

References

1. Brown DM, Kaiser PK, Michels M et al (2006) Ranibizumab versus verteporfin for neovascular age-related macular degeneration. N Engl J Med 355:1432–1444. https://doi.org/10.1056/NEJMoa062655
2. Stasch-Bouws J, Eller-Woywod SM, Schmickler S et al (2020) IVOM quality assurance in Westfalen-Lippe : Structure of quality assurance and results of the pilot study Q-VERA. Ophthalmologe 117:336–342. https://doi.org/10.1007/s00347-019-01030-3
3. Brinkmann CK, Chang P, Schick T et al (2019) Baseline diagnostics and initial treatment decision for anti-vascular endothelial growth factor treatment in retinal diseases : Comparison between results by study physician and reading centers (ORCA/OCEAN study). Ophthalmologe 116:753–765. https://doi.org/10.1007/s00347-018-0805-y
4. Gulshan V, Peng L, Coram M et al (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 316:2402. https://doi.org/10.1001/jama.2016.17216
5. Lee CS, Tyring AJ, Deruyter NP et al (2017) Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomed Opt Express 8:3440–3448. https://doi.org/10.1364/BOE.8.003440
6. Burlina PM, Joshi N, Pekala M et al (2017) Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. JAMA Ophthalmol 135:1170–1176. https://doi.org/10.1001/jamaophthalmol.2017.3782
7. Abràmoff MD, Lavin PT, Birch M et al (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 1:39. https://doi.org/10.1038/s41746-018-0040-6
8. De Fauw J, Ledsam JR, Romera-Paredes B et al (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24:1342–1350. https://doi.org/10.1038/s41591-018-0107-6
9. Russakoff DB, Lamin A, Oakley JD et al (2019) Deep Learning for Prediction of AMD Progression: A Pilot Study. Invest Ophthalmol Vis Sci 60:712. https://doi.org/10.1167/iovs.18-25325
10. Wang Y, Tu D, Du J et al (2019) Classification of Subcortical Vascular Cognitive Impairment Using Single MRI Sequence and Deep Learning Convolutional Neural Networks. Front Neurosci 13:627. https://doi.org/10.3389/fnins.2019.00627
11. Maetschke S, Antony B, Ishikawa H et al (2019) A feature agnostic approach for glaucoma detection in OCT volumes. PLoS ONE 14:e0219126. https://doi.org/10.1371/journal.pone.0219126
12. Li Z, Guo C, Nie D et al (2020) Deep learning for detecting retinal detachment and discerning macular status using ultra-widefield fundus images. Commun Biol 3:15. https://doi.org/10.1038/s42003-019-0730-x
13. Kermany DS, Goldbaum M, Cai W et al (2018) Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
14. Liefers B, Colijn JM, González-Gonzalo C et al (2020) A Deep Learning Model for Segmentation of Geographic Atrophy to Study Its Long-Term Natural History. Ophthalmology 127:1086–1096. https://doi.org/10.1016/j.ophtha.2020.02.009
15. Schmidt-Erfurth U, Waldstein SM, Klimscha S et al (2018) Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. Invest Ophthalmol Vis Sci 59:3199–3208. https://doi.org/10.1167/iovs.18-24106
16. Yim J, Chopra R, Spitz T et al (2020) Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med 26:892–899. https://doi.org/10.1038/s41591-020-0867-7
17. Lee CS, Baughman DM, Lee AY (2016) Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. Ophthalmol Retina 1:322–327. https://doi.org/10.1016/j.oret.2016.12.009
18. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A et al (2018) Machine Learning to Analyze the Prognostic Value of Current Imaging Biomarkers in Neovascular Age-Related Macular Degeneration. Ophthalmol Retina 2:24–30. https://doi.org/10.1016/j.oret.2017.03.015
19. Rothaus K, Farecki M-L, Mussinghoff P et al (2019) Analysis of the "Portal" Care Model - Examination of the Outcome Quality of IVOM Therapy with Regard to Latency Periods in Exudative AMD. Klin Monbl Augenheilkd 238:293–301. https://doi.org/10.1055/a-0982-5294
20. IVAN Study Investigators, Chakravarthy U, Harding SP et al (2012) Ranibizumab versus bevacizumab to treat neovascular age-related macular degeneration: one-year findings from the IVAN randomized trial. Ophthalmology 119:1399–1411. https://doi.org/10.1016/j.ophtha.2012.04.015
21. CATT Research Group, Martin DF, Maguire MG et al (2011) Ranibizumab and bevacizumab for neovascular age-related macular degeneration. N Engl J Med 364:1897–1908. https://doi.org/10.1056/NEJMoa1102673

22. Mazzaferri J, Beaton L, Hounye G et al (2017) Open-source algorithm for automatic choroid segmentation of OCT volume reconstructions. Sci Rep 7:42112. https://doi.org/10.1038/srep42112
23. Bradski G (2000) The OpenCV Library: Resampling using pixel area relation. https://docs.opencv.org/3.4/da/d54/group__imgproc__transform.html. Accessed 17 Oct 2021
24. Lim G, Bellemo V, Xie Y et al (2020) Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. Eye Vis Lond Engl 7:21. https://doi.org/10.1186/s40662-020-00182-7
25. Dozat T (2016) Incorporating Nesterov Momentum into Adam. https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ. Accessed 17 Oct 2021
26. Chollet F (2015) Keras. https://keras.io. Accessed 17 Oct 2021
27. Abadi M, Agarwal A, Barham P, Brevdo, E (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org. Accessed 17 Oct 2021
28. Simonyan K, Vedaldi, Andrea, Zisserman, Andrew (2014) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034. Accessed 17 Oct 2021
29. Kotikalapudi R (2017) What is Saliency? https://raghakot.github.io/keras-vis/visualizations/saliency/#what-is-saliency. Accessed 17 Oct 2021
30. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577
31. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382. https://doi.org/10.1037/h0031619
32. Mishra Z, Ganegoda A, Selicha J et al (2020) Automated Retinal Layer Segmentation Using Graph-based Algorithm Incorporating Deep-learning-derived Information. Sci Rep 10:9541. https://doi.org/10.1038/s41598-020-66355-5
33. Maloca PM, Lee AY, de Carvalho ER et al (2019) Validation of automated artificial intelligence segmentation of optical coherence tomography images. PLoS ONE 14:e0220063. https://doi.org/10.1371/journal.pone.0220063
34. Kurmann T, Yu S, Márquez-Neila P et al (2019) Expert-level Automated Biomarker Identification in Optical Coherence Tomography Scans. Sci Rep 9:13605. https://doi.org/10.1038/s41598-019-49740-7
35. Schlegl T, Waldstein SM, Bogunovic H et al (2018) Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. Ophthalmology 125:549–558. https://doi.org/10.1016/j.ophtha.2017.10.031
36. Yan Q, Weeks DE, Xin H et al (2020) Deep-learning-based Prediction of Late Age-Related Macular Degeneration Progression. Nat Mach Intell 2:141–150. https://doi.org/10.1038/s42256-020-0154-9