End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging

David Romo-Bucheli, Ursula Schmidt-Erfurth, and Hrvoje Bogunović

***Abstract*—Neovascular age-related macular degeneration (nAMD) is nowadays successfully treated with anti-VEGF substances, but inter-individual treatment requirements are vastly heterogeneous and currently poorly plannable resulting in suboptimal treatment frequency. Optical coherence tomography (OCT) with its 3D high-resolution imaging serves as a companion diagnostic to anti-VEGF therapy. This creates a need for building predictive models using automated image analysis of OCT scans acquired during the treatment initiation phase. We propose such a model based on deep learning (DL) architecture, comprised of a densely connected neural network (DenseNet) and a recurrent neural network (RNN), trainable end-to-end. The method starts by sampling several 2D-images from an OCT volume to obtain a lower-dimensional OCT representation. At the core of the predictive model, the DenseNet learns useful retinal spatial features while the RNN integrates information from different time points. The introduced model was evaluated on the prediction of anti-VEGF treatment requirements in nAMD patients treated under a pro-re-nata (PRN) regimen. The DL model was trained on 281 patients and evaluated on a hold-out test set of 69 patient. The predictive model achieved a concordance index of 0.7 in regressing the number of received treatments, while in a classification task it obtained an 0.85 (0.81) AUC in detecting the patients with low (high) treatment requirements. The proposed model outperformed previous machine learning strategies that relied on a set of spatio-temporal image features, showing that the proposed DL architecture successfully learned to extract the relevant spatio-temporal patterns directly from raw longitudinal OCT images.**

***Index Terms*—optical coherence tomography, deep learning, age-related macular degeneration, longitudinal imaging**

I. Introduction

TREATMENT based on anti-Vascular Endothelial Growth Factor (anti-VEGF) substances has been shown to be very effective and to significantly improve visual acuity outcomes in patients with neovascular AMD (nAMD) [1], [2], a leading cause of severe vision loss in the elderly population in developed countries [3]. Nevertheless, anti-VEGF therapy is expensive and requires frequent and long-term follow-up, which is currently poorly plannable as recurrent and persistent neovascular exudation demonstrates a huge inter-individual variability.

Optical coherence tomography (OCT) has become a standard of care in ophthalmology [4]. It is the most commonly used imaging modality in ophthalmology with over 30 million scans being performed every year in US alone [5]. It had a pivotal role in the development of antiangiogenic therapies for the treatment of nAMD because it is able to visualize the macular fluid and is hence used as a diagnostic companion to anti-VEGF treatment [6]. The use of OCT as a "VEGF meter" in guidance of therapy has already resulted in public savings of more than $10 billion [5]. In comparison to fluorescein angiography (FA), OCT imaging is a fast, safe, noninvasive technique that complemented FA imaging by providing cross-sectional images of the retina.

Intravitreal treatment decisions are hence currently made based on high-resolution 3D OCT scans taken at the time of continued visits [7]. This decision is predominantly driven by two criteria: (a) the presence of disease activity (retinal fluid) and (b) a perceived loss in visual acuity. Specifically, the disease activity is assessed by examination of the patient's retinal morphology via OCT imaging. In randomized clinical trials, the pro-re-nata (PRN) regimen, i.e. treatment is given when needed, and the treat-and-extend (TE) regimen extending intervals until fluid recurs have been shown to work well, with outcomes comparable to the more intensive monthly regimen [7], [8].

Real-world reports unfortunately conclude that it is difficult for patients and/or physicians to strictly adhere to a rigorous follow-up schedule, resulting in patients receiving fewer injections and worse visual outcomes when compared with prospective clinical trials [9], [10]. Notably, patients with a better visual acuity at the beginning of the treatment are vulnerable to vision loss. This dilemma highlights the need for additional image-guided predictive tools aiming at managing anti-VEGF treatment in an optimal manner. Particularly, the availability of artificial intelligence for predictive modeling is expected to help at the same time minimize the number of visits while supporting maximal visual function recovery. Such models will allow the practicing clinicians to adequately adjust the scheduling of patient visits, leading to optimal use of resources, in a hope of improving real-world treatment outcomes.

A comprehensive long-term management would be based on a predictive model, which utilizes all the available data from prior OCT scans in patient's medical history. Typically, all the relevant clinical information within an observation window (containing more than one observation time-point) is known. This information could then be used to predict the clinical outcomes within a future prediction window. However,

D. Romo-Bucheli, U. Schmidt-Erfurth and H. Bogunović are with the Laboratory for Ophthalmic Image Analysis, Department of Ophthalmology and Optometry, Medical University Vienna, Austria. (email: hrvoje.bogunovic@meduniwien.ac.at, ursula.schmidt-erfurth@meduniwien.ac.at)

creating such a predictive model from longitudinal OCT data raises several challenges. First, each OCT volume has a large dimensionality (a 3D OCT volume typically contains approx. 70 Mio voxels). Second, though the total number of OCTs may be large, longitudinal OCT usually is available for only a relatively small number of patients (in $100s$). Finally, there is currently a lack of pretrained deep learning models that would enable transfer learning in the setting of prediction from longitudinal medical data.

A. Related Work

Recent studies on large populations have shown that deep learning (DL) approaches are able to correctly identify common retinal diseases from OCT images. In [11], a total of $207,130$ OCT B-scans images were used to train a DL model to distinguish four different retinal categories. In a final evaluation, the classes were grouped into "urgent referral" and "non-urgent", and in this classification task the DL model yielded a 99.9% area under the receiver operating characteristic curve (AUC). Similarly, in [12] a set of $14,884$ OCT volumes was used to train a two-stage DL framework. In the first stage, deep neural network segments the retinal tissue into up to 15 different classes related to the retinal anatomy and pathology, and image artifacts. Then the derived segmentation maps are supplied into a second stage network that is trained to perform differential diagnosis and provide referral suggestions. The DL framework achieved expert-like performance in the identification and referral of sight-threatening retinal diseases with a 5.5% error rate.

Deep learning from retinal OCT has also been employed for predictive modeling of future outcomes or natural disease progression. In [13] they predicted a response to anti-VEGF treatments in patients suffering from diabetic macular edema (DME). However, their prediction was performed from a single pretreatment OCT, and did not integrate longitudinal information from the treatment initiation phase. Similarly, in [14], they developed a predictive model of conversion from intermediate to neovascular AMD. The proposed AMDnet takes also a single OCT, and they train the model per individual B-scan, where the prediction value at the OCT-level was obtained by taking the mean of each volume's individual B-scan predictions.

Techniques based on DL have recently been introduced for prediction tasks from longitudinal (several time points) medical data in general. In [15], the patient electronic health records were integrated across different time-points by using several DL approaches including recurrent neural networks (RNN) and convolutional neural networks (CNN). The authors reported that this integration improves the performance in the prediction task of cardiovascular events. Likewise, in [16], heterogeneous medical data from several time-points is integrated by using a long short-term memory RNN (LSTM) to predict Alzheimer's disease progression with a 99% accuracy. In a recent work [17], a classification framework for Alzheimer's disease diagnosis from longitudinal magnetic resonance imaging is proposed. Reported experimental results showed a 91.33% accuracy in distinguishing Alzheimer patients from normal controls.

In longitudinal retinal OCTs of nAMD patients the following related works stand out. In a study of patients receiving anti-VEGF treatment [18], traditional machine learning (ML) techniques were used to analyze the prognostic value of automatically extracted sets of retinal biomarkers. The study aimed to predict the final visual acuity in AMD patients based on initial treatment responses. Similarly, in [19], they predicted macular edema recurrence after anti-VEGF therapy in patients with retinal vein occlusion (RVO) from longitudinal retinal OCT. However, they do not use a deep learning model but rather rely on retinal segmentation and the resulting series of 2D retinal thickness maps were used to train a classifier based on logistic regression with L1 regularization.

In another longitudinal study, the closest related work, ML techniques were used to identify patients with high/low treatment requirement in 317 nAMD patients [20]. The longitudinal predictions are done by first computing $2D$ en-face feature maps (retinal layers thickness, retinal fluid volume/area, among others) for each OCT volume. These feature maps are then summarized into spatio-temporal feature vectors. The trained ML model yielded an 0.77 (0.70) AUC for identifying patients with high/low anti-VEGF treatment requirement. Thus the current state of the art in longitudinal retinal OCT is based on traditional ML approaches as the extracted spatio-temporal features serve as an effective dimensionality reduction. However, these tasks are inherently limited to the features extracted from OCT biomarkers that are clinically predefined and can be segmented while there may be other predictive image patterns that are currently overlooked, but could be found and exploited using an end-to-end deep learning approach.

Contributions: The main contribution of this paper is the design and validation of an end-to-end DL methodology for prediction from longitudinal retinal OCT applied to predicting treatment requirements for the management of nAMD. The proposed methodology integrates imaging information across several OCT volumes and to the authors' knowledge, this is the first method to use an end-to-end DL set-up for a longitudinal OCT prediction task. A second contribution of our work, is the design of a pre-processing pipeline aiming to reduce the OCT volume dimensionality, by using a set of star-shaped sampling planes, allowing the model to learn predictive patterns in setting with limited amount of data. A third contribution, is the interrogation of the deep learning model to identify novel spatio-temporal imaging patterns, especially relevant in the tasks where currently there are no clinically known predictive imaging biomarkers. Our results show that the proposed methodology was able to deal with the constraints imposed by the longitudinal OCT data, and to effectively integrate the spatio-temporal information for the clinically relevant task of guiding nAMD treatment requirements based on solid personalized prediction.

II. Materials and Methods

The longitudinal OCT retreatment prediction task consists of a set of acquired $3D$ scans in the *observation window* from which a prediction within the future *prediction window* is made. In this section, we describe first the longitudinal

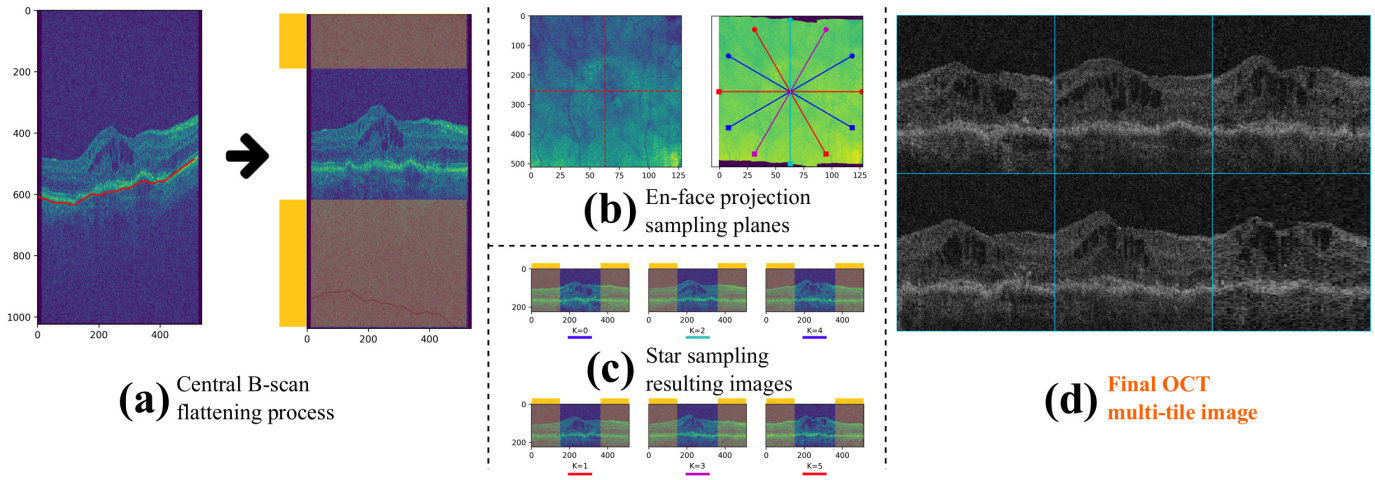

Fig. 1. Schematic diagram of the OCT pre-processing pipeline. (**a**) Flattening of the B-scan using Bruch's membrane (BM), shown as a red line, as reference. Discarded image regions are shown in yellow. (**b**) Star sampling: The OCT scan center is the intersecting point of the sampling planes. (**c**) Images corresponding to k sampling planes with discarded regions shown in yellow. (**d**) Resulting single multi-tile image. In the diagram, only $k = 6$ sampling planes are used to illustrate the pipeline.

OCT dataset followed by the experimental set-up used in this work. We then describe the pre-processing algorithms used for reducing the dimensionality of the OCT volume. Further, we present the deep learning architecture used to address the automatic prediction of anti-VEGF treatment needs and the deep learning training setting. Finally, we describe a feature/biomarker based prediction strategy used as baseline.

A. Dataset description and curation

Data of treatment-naive nAMD patients undergoing a PRN protocol with monthly visits over two (2) years is used. A total of 423 patients with available treatment data (time-point injection decisions) and associated initiation phase OCT imaging were included in the present study. The effective number of injections (n_{inj}) received by the patients was used to define a set of treatment requirement categories: High ($n_{inj} \geq 16$), intermediate ($5 < n_{inj} < 16$) and low ($n_{inj} \leq 5$), where high and low categories correspond to the first and third quartile of the population, respectively. The defined treatment requirement category is the target in our classification task. The study complies with the Declaration of Helsinki, and the Ethics Committee at the Medical University of Vienna approved the use of this data in post hoc analysis.

Curated dataset: Clinically, the decision not-to-inject in patients that exhibit disease activity puts their vision at risk. Hence controlling for such investigator decisions is needed for our task, since it is not uncommon for small amounts of fluid to be neglected in a routine examination[21]. For doing so, we ran a deep learning based automated quantification fluid algorithm [22] on the whole set of OCT images associated with non-injection events during the 2 year study. In an independent previous evaluation, this quantification algorithm yielded a performance of 0.93 AUC in the retinal fluid detection task [22]. After using this algorithm, we computed the median fluid volume ($q2$) and the interquartile range (IQR) for all the non-injection events in the dataset. Those non-injection events in which the automatically computed fluid volume was larger than $q2 + 1.5 \times IQR$ were regarded as disagreements. Patients that had more than 3 disagreements in the entire 2 year treatment period were excluded from the curated dataset. After this procedure the total set of patients used in this study resulted in 350 individuals.

B. Evaluation of the prediction of retreatment requirement

We defined three different prediction tasks based on the treatment requirement information as follows.

Treatment requirement regression: For this task we tried to predict directly the treatment requirement score (RQS) defined as the number of received injections divided by the total number of visits not counting the standardized initiation phase of three consecutive monthly injections. This number ranges between 0 (for a patient that did not require any additional injections during the subsequent PRN phase) and 1 (for a patient requiring injections on a monthly basis). To evaluate the performance on this regression task we used the following metrics: R-squared, the Pearson correlation coefficient, and the concordance index.

Multi-class classification task: In this task, we defined the high, intermediate and low treatment requirement categories as defined in the Subsection II-A. We evaluated the performance on this classification task by computing the confusion matrix and the overall accuracy. The accuracy, specificity and sensitivity (recall) per class were also computed.

Binary classification tasks: This task corresponds to the identification of either low or high treatment requirement patients. This particular problem has been addressed in a previous work [20]. Specifically, both problems can be defined as two separate binary classification tasks: high (low) group vs. remaining patients. To evaluate the performance in this task, we used the area under curve (AUC) of the receiver operating characteristic (ROC).

C. OCT volume pre-processing

The OCT scan pre-processing starts by using a movement correction process to reduce the misalignment across OCT B-scans following the algorithm described in [23]. Afterwards, Bruch's membrane (BM) is automatically detected in the OCT volume by using the Iowa reference algorithms [24], [25]. The curve corresponding to the BM segmentation is used to flatten each B-scan to a predefined level. This operation reduces the variability across OCT volumes and facilitates the subsequent pre-processing. Imaging data corresponding to 658 μm and 209 μm above/under the BM delineation is then cropped from the B-scans to obtain a region with a height of 224 pixels centered on the retina. After this operation, a star pattern sampling is applied to the volume yielding a set of images, each corresponding to one sampling plane. The sampling is carried out by using the en-face center of the scan as the intersecting point of the k sampling planes (Fig. 1). Afterwards the k obtained images are cropped while preserving the information within $\sim$ 3 mm of the central region of the retina, which is considered to be most relevant to central vision function. The resulting images are finally arranged into a 2D grid (Fig. 1(d)) that contains information from an individual patient at a specific time-point.

D. Deep learning architecture

The DL approach developed in this work, is comprised of three components (Fig. 2):

DenseNet architecture: The densely connected network (DenseNet) [26] provides the feature extraction process in the OCT images. The DenseNet architecture consist of multiple "dense blocks", or sets of convolutional layers (with kernel size 1×1 and 3×3). The convolutional layers within a dense-block are connected in such a way that their inputs correspond to ALL the outputs of the preceding layers. The selection of this architecture was motivated by the relatively low number of parameters, when compared with other standard architectures, and additionally it has been empirically observed that DenseNet has good convergence properties[27].

Recurrent neural network: A standard RNN was used to integrate the information of the OCT images across multiple time-points of the initiation phase. The RNN architecture corresponds to a two-layer Elman RNN with hyperbolic tangent (tanh) activation function [28]. In a nutshell, the recurrent neural algorithm is a fully connected layer that has a feedback connection which allows temporal dynamic changes to be modeled.

Fully connected layer: This layer integrates the spatio-temporal information from the RNN and generates a category prediction. This component consist of a linear layer followed by a non-linear activation function. The output correspond to the patient's probability of being part of each treatment requirement category.

The whole DL strategy was implemented using the pytorch library [29]. The pytorch library's dynamic graph computation functionality allowed us to implement a seamless transition between the DenseNet and the RNN for the gradient computation in the backpropagation algorithm (weight optimization process) in an end-to-end training process from scratch, i.e., no pre-trained weights were used to fine-tune the model parameters.

At training stage, the cross entropy (CE) loss was used for the classification tasks as it models the error between two probability distributions in the framework of *maximum likelihood* estimation. For the regression task the $L1$ loss was used. The L1 loss penalizes the $L1$ norm between the predicted and actual values of the inputs:

$$loss(x) = |x_n - y_n| \tag{1}$$

Where x_n corresponds to the predicted value and y_n the ground truth for a specific sample.

On the other hand, the CE loss is associated to a n dimensional output vector, each dimension associated to the probability of a particular input being categorized in one of the n classes. Then, the associated loss for each class can be described as follows:

$$loss(x, k_{\text{class}}) = -x[k_{\text{class}}] + log(\sum_j exp(x[j])) + \lambda \parallel w \parallel_2^2 \tag{2}$$

Where $x[\text{k}_{\text{class}}]$ corresponds to the probability for the correct label, given a specific example, and $x[j]$ corresponds to the probability assigned to class j. The final loss is computed by aggregating all the class losses. The L_2 weight penalization term was also added to regularize the network weights w and reduce the likelihood of DL model over-fitting. Additionally, data augmentation techniques were applied: OCT volume flipping (left/right horizontal flipping) and image intensity augmentation: brightness and contrast [30].

E. Deep learning training setup

Evaluation of the presented methodology was performed on a randomly fixed set-up. This particular setup is widely used in the DL research community due to the computationally expensive DL training process. Additionally, the longitudinal OCT prediction task requires a relatively large number of OCT images and the DL model selection and evaluation would become computationally prohibitive if a more exhaustive set-up such as cross-validation is used. The available data was split into train, validation and test sets ($\sim 70\%$ - 247 patients, $\sim 10\%$ - 34 patients, $\sim 20\%$ - 69 patients) using random sampling stratified by the treatment requirement category (high, intermediate, and low). In the pre-processing step, the number of sampling planes was set to $k = 16$. Accordingly, the resulting sampled images for each OCT volume were arranged in a 4×4 multi-tile image.

The DenseNet architecture was comprised of (n_{DB}) dense-blocks with a growth rate of 12, and each dense-block was composed by n_L convolutional layers. Drop-out layers were also included in the dense-block and the drop-out rate for all of them was set to 0.4. The recurrent neural network is fed with the features extracted from the DenseNet architecture and its output is fed to a fully connected layer to finally obtain a 3-dimensional category vector (classification task), or a 1-dimensional output (regression task).

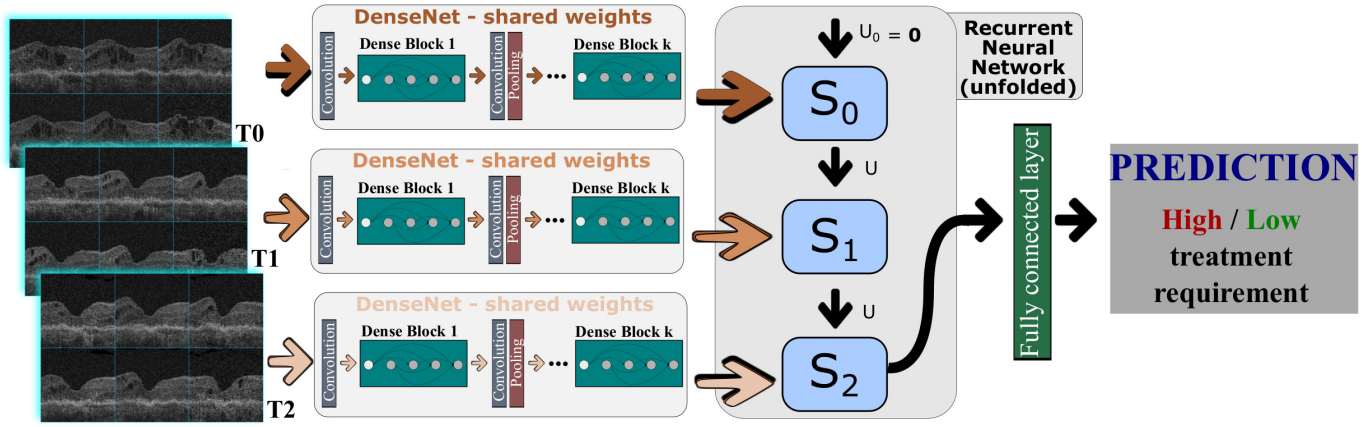

Fig. 2. Deep learning architecture used in this paper. After pre-processing the initiation phase OCT volumes of a specific patient ($T0$: day 0, $T1$: day 30, and $T2$: day 60), three different multi-tile images are obtained. These images are fed into the densely connected network (mainly composed by dense-blocks, convolutional and pooling layers). The extracted features of each multi-tile image are integrated using a recurrent neural network (RNN). The unfolded representation is presented in the scheme with the initial state U_0 equal to a vector of zeros. Finally, a fully connected layer predicts the treatment requirement category: High, intermediate or low.

Regression task: In this case, the L1 loss was optimized using the SGD algorithm and a fixed number of epochs ($n_{epochs} = 300$). The learning rate was set to $l_{rate} = 1e^{-3}$ and an automatic learning schedule was used [1]. Six different configuration models varying the number of dense-blocks ($n_{DB} = 5$ or $n_{DB} = 6$) and number of convolutional layers. After each training epoch, the L1 loss was computed to evaluate the performance in the validation set and the model associated with the best running performance was stored. We stored 6 different models, with different parameters regarding the number of dense-blocks ($n_{DB} = 5$ or $n_{DB} = 6$) and number of convolutional layers $\{n_L = 2, n_L = 3$ or $n_L = 4\}$ within each dense block. The resulting models from the six (6) different configurations were evaluated in the validation set and the best performing model was selected.

Classification tasks: A similar procedure, to that described for regression task, was used for the classification task. The main difference is that here the CE loss, instead of the L1 loss, was minimized using the SGD algorithm. The same learning rate schedule with an initial learning rate ($l_{rate} = 1e^{-3}$) was used. In this case, again 6 different models, with varying parameters as described in the previous paragraph, were trained and the best performing model was selected. The results on the validation set used to select the parameter configuration are presented in Table I. The geometric accuracy, defined as the geometric average of the recall per-class, was used as the selection metric. After comparing the performance, the model with $n_L = 4$ and $n_{DB} = 5$ was selected.

F. Feature/biomarker based prediction strategy

An additional prediction strategy based on a random forest classifier was implemented for comparison. The prediction approach replicates the methodology proposed in [20] to predict treatment needs in anti-VEGF therapy. A set of quantitative spatio-temporal OCT features, which are computed from automated segmentation of retinal fluid and retinal layers, are used to feed the random forest classifier. The automated retinal fluid segmentations were obtained by using a deep learning segmentation model trained on a different set of OCT scans [22]. Respectively, the layer segmentation was carried out by using the Iowa reference algorithms[24], [25].

[1]http://blog.dlib.net/2018/02/automatic-learning-rate-scheduling-that.html

TABLE I
GEOMETRIC ACCURACY METRIC FOR THE MODEL SELECTION PROCESS

N. of Dense Blocks (DB)	N. of conv. layers within the DB	Geometric accuracy
5	2	0.45
6	2	0.51
5	3	0.53
6	3	0.4
5	*4*	*0.59*
6	4	0.55

III. RESULTS

The DL model performance in the retreatment prediction task was evaluated on the held-out test set. Each task (regression, two-class, and three-class) was trained and evaluated separately.

A. Regression task

The results for the regression task are presented in Fig. 4. A Pearson correlation coefficient $R = 0.59$, and coefficient of determination $R^2 = 0.22$ were obtained by the selected regression model. Additionally, a concordance index of 0.7 was achieved, i.e., 70% of patient-pairs were concordant. This indicates that when randomly choosing any pair of predicted RQS and sorting them, there is a 70% probability that the true RQS order is identical.

B. Classification tasks

a) Three-class: When evaluating the model on the full three-class classification problem, we obtained the confusion

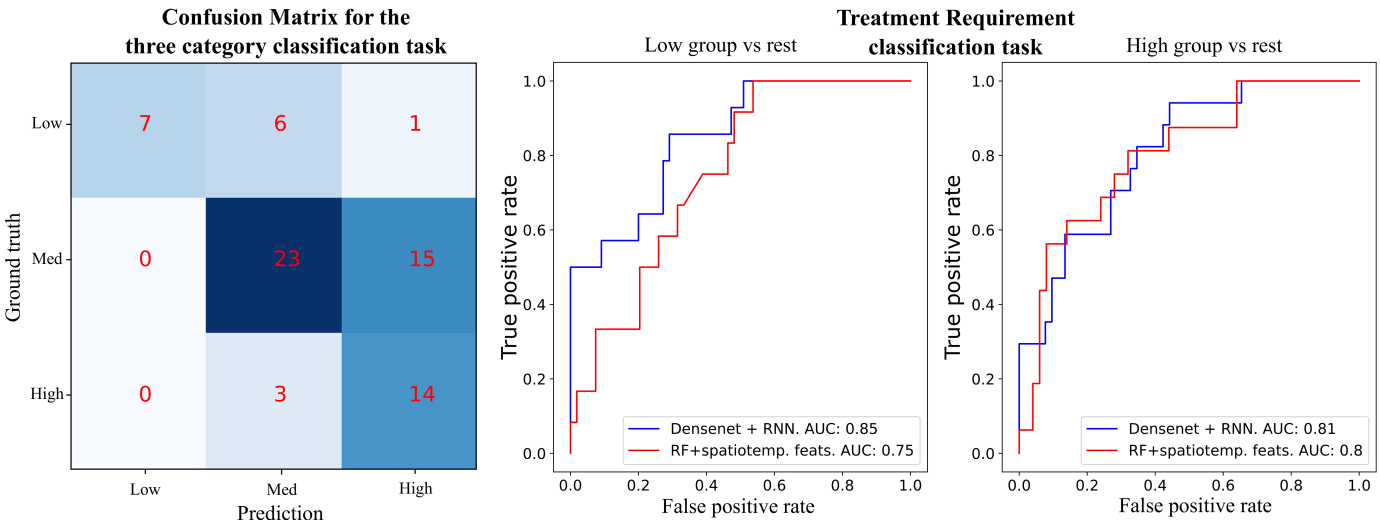

Fig. 3. Receiver operating characteristic (ROC) curves for the classification tasks evaluated in this paper. *Left:* Confusion matrix for the full three category classification task: high, intermediate (med), and low. In this case, we used the actual output selection process based on softmax learned by the DL model. *Center:* Low treatment requirement group vs remaining patients. *Right:* High treatment requirement group vs remaining patients. For both binary classification tasks, the presented methodology (blue line) performs equally or better than a stratification strategy based on automated segmentation of several retinal structures(red).

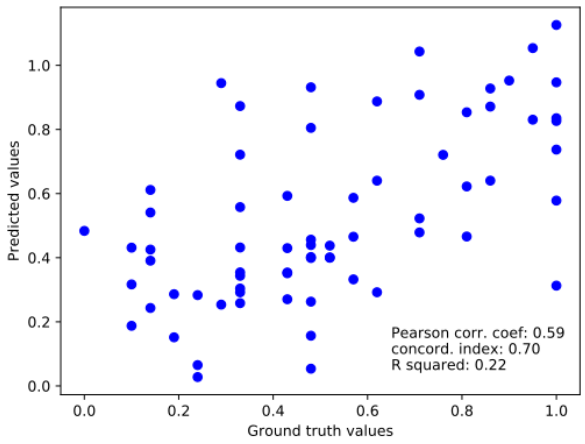

Fig. 4. Ground truth and predicted treatment requirement score (RQS) values on the held out test set.

matrix in the leftmost column of Fig. 3. There was considerable overlap between the low and intermediate groups. Likewise, there was considerable overlap between the high and intermediate groups. This fact is reflected on the overall accuracy metric (0.64) for the three-class classification task. Additionally, the accuracy, specificity, and sensitivity (recall) per class are shown in Table II. The per-class results indicate that the model was able to identify accurately and with high specificity, a large number of retinas with high treatment needs. On the high class, the accuracy was worse, indicating that there was considerable overlap with the intermediate group. Finally, the intermediate group has lower accuracy and sensitivity. This result is reasonable, taking into account that this particular group corresponds to the most challenging for predicting response, from the clinical point of view.

TABLE II
ACCURACY, SENSITIVITY AND SPECIFICITY METRICS PER CLASS FOR THE THREE-CLASS PROBLEM.

Class	Accuracy	Sensitivity	Specificity
Low	0.9	0.5	1.0
Intermediate	0.65	0.61	0.71
High	0.72	0.82	0.69

b) Two-class: The results for the two different binary classification task are summarized in the two rightmost columns of Fig. 3. The presented DL model outperforms the baseline in both binary classification tasks: low (high) treatment requirement group vs remaining patients. We additionally, applied the Delong significance test to determine if the ROC curves between the proposed approach and the baseline are significantly different [31]. The DL model yielded an AUC of 0.81 in the high vs remaining classification task outperforming the baseline model (AUC = 0.8). However the improvement was not statistically significant (p-value = 0.702). This implies that relevant predictive patterns are already covered by the established fluid volume features of the baseline model. Meanwhile, for the low vs. remaining cases classification task the DL model yielded an AUC of 0.85 substantially outperforming the baseline model (AUC = 0.75), a substantial improvement that was close to being statistically significant (p-value = 0.084). This implies that for this class the model was able to learn new predictive patterns going beyond the established fluid volume features included by the baseline model. This was expected as retinas requiring low amount of treatments are mostly dry and fluid-based features are not sufficient to identify them

C. Interpretation of the model decisions

The DL model performance was similar (for the high vs. rest) or better (low vs. rest) to that obtained by a strategy based on automated segmentation of diverse retinal structures. However, a pervasive problem with DL models is that they produce black-box models and their interpretation is often challenging. Thus, several visualization techniques have been developed to address this shortcoming. These techniques aim to determine regions of the image that contribute heavily to the prediction of a DL model to confirm clinical robustness.

In this work, we adapted the occlusion sensitivity method proposed in [32] to visualize the predictions of our DL model in the longitudinal OCT prediction task. The method consists in modifying the input images by setting the intensity values within a rectangular patch to a pre-defined value, i.e, 0 would define a black patch. Then, this patch is translated across the whole image to generate a series of perturbed images. The model is supplied with the perturbed images and the resulting predictions are stored. The probability drop with respect to the original image prediction is then used to build a normalized attribution heatmap. For this task, a black patch of 32×32 was used. Additionally, the perturbations across the multi-tile images in the 3 different time points were synchronized, i.e., the black patch was generated within the same coordinates for the three images used on the prediction task: $T0$, $T1$ and $T2$.

Examples of the generated heatmaps per patient and the corresponding multi-tile images are presented in Fig. 5. A common finding in retinas with high treatment requirements (top row) is a strong deformation at $T0$ (with the presence of retinal fluid - black regions). It is not unusual to observe small areas with retinal fluid even at $T2$ (2 months after the initial treatment). On the other hand, retinas with low and intermediate requirement needs, also present some fluid in the initial time point, albeit the retinal deformation is smaller in $T0$ and almost negligible in $T1$ and $T2$ in comparison to the high treatment retinas.

Finally, the normalized attribution heatmaps of the retinas were used to obtain a "representative" attribution heatmap per treatment requirement category. Such representative heatmaps were computed by averaging the normalized heatmaps of all the *correctly* classified retinas in each of the three category classification tasks. The resulting "representative" maps for the high, intermediate and low treatment requirement cases are shown in Fig. 6 with high (low) attribution values associated to red (blue) regions in the multi-tile image. We can observe that while for the high and the intermediate requirement groups the heatmap is scattered across the whole image grid, the average heatmap for the low group is focused on a small region below Bruch's membrane. Furthermore, the average representation of the high and low categories seem to be complementary, i.e, red regions in the low group correspond to blue regions in the high group. This might indicate that the learned evaluation function might actually strongly associate certain regions of the multi-tile image to either the high/low class, while assigning a much more homogeneous attribution map to the intermediate class.

IV. Discussion

An end-to-end DL based approach for anti-VEGF treatment requirement prediction on longitudinal OCT images was presented. Such longitudinal OCT prediction tasks are challenging to address with DL methods because there is only a limited number of training samples available (100s of patients), while OCT volumes per se exhibit extremely high dimensionality (millions of anisotropic voxels), the so called "large p, small n" problem. We partially addressed this problem by using a pre-processing pipeline that proved effective in reducing the OCT dimensionality while still retaining relevant clinical information (i.e., associated to the central region of the fovea).

The treatment requirement prediction herein addressed was posed as a **(a)** regression task, **(b)** three-category classification task, and **(c)** two binary classification tasks. The decision to cast the longitudinal OCT prediction task in these three different settings allowed us to evaluate the ability of our DL framework to address the prediction problem with different stages of "difficulty".

The regression of RQS values is the hardest stage of the problem. Though we found that the DL framework was able to generate estimations correlated with the actual RQS ground truth values, we also found that there was substantial variability between the predictions and the ground-truth across all the RQS values (Fig 4). However, in this setting it is not clear, if a particular RQS interval is harder or easier to predict than the others. This difficulty is addressed in the three category classification problem, in which we split the prediction interval into three different groups. We found out that the model was better in identifying the extreme levels associated with the high and low groups. However, there was a considerable amount of cases misclassified in the intermediate group. This resulted in a low accuracy score for the three-class prediction task. Finally, in the two binary classification tasks the DL separately focused on the two groups that are clinically most relevant as they are prone to progressive visual loss: patients with high (low) treatment needs. The results showed that the performance for predicting high requirements was similar (slightly better) to that yielded by the baseline feature-based method. This might indicate that the DL model is probably capturing information highly correlated to the information used by the baseline model, i.e. the intraretinal and subretinal fluid which is present for longer periods in this subgroup. Nevertheless, a more interesting result was observed for predicting the low requirements, in which the DL model had a noticeably better performance. That subgroup of patients is characterized by very little retinal fluid present and hence the traditional biomarker does not describe the retina in sufficient detail to distinguish well this subgroup. This is precisely where DL models can take advantage of their automated feature learning and our result indicates that the DL model is capturing additional information that is not being utilized by the baseline ML method.

The above finding/observation is further supported by our occlusion sensitivity analysis (Section III-C). In the "representative" attribution heatmaps obtained for each category, some patterns could be observed. For the high and intermediate

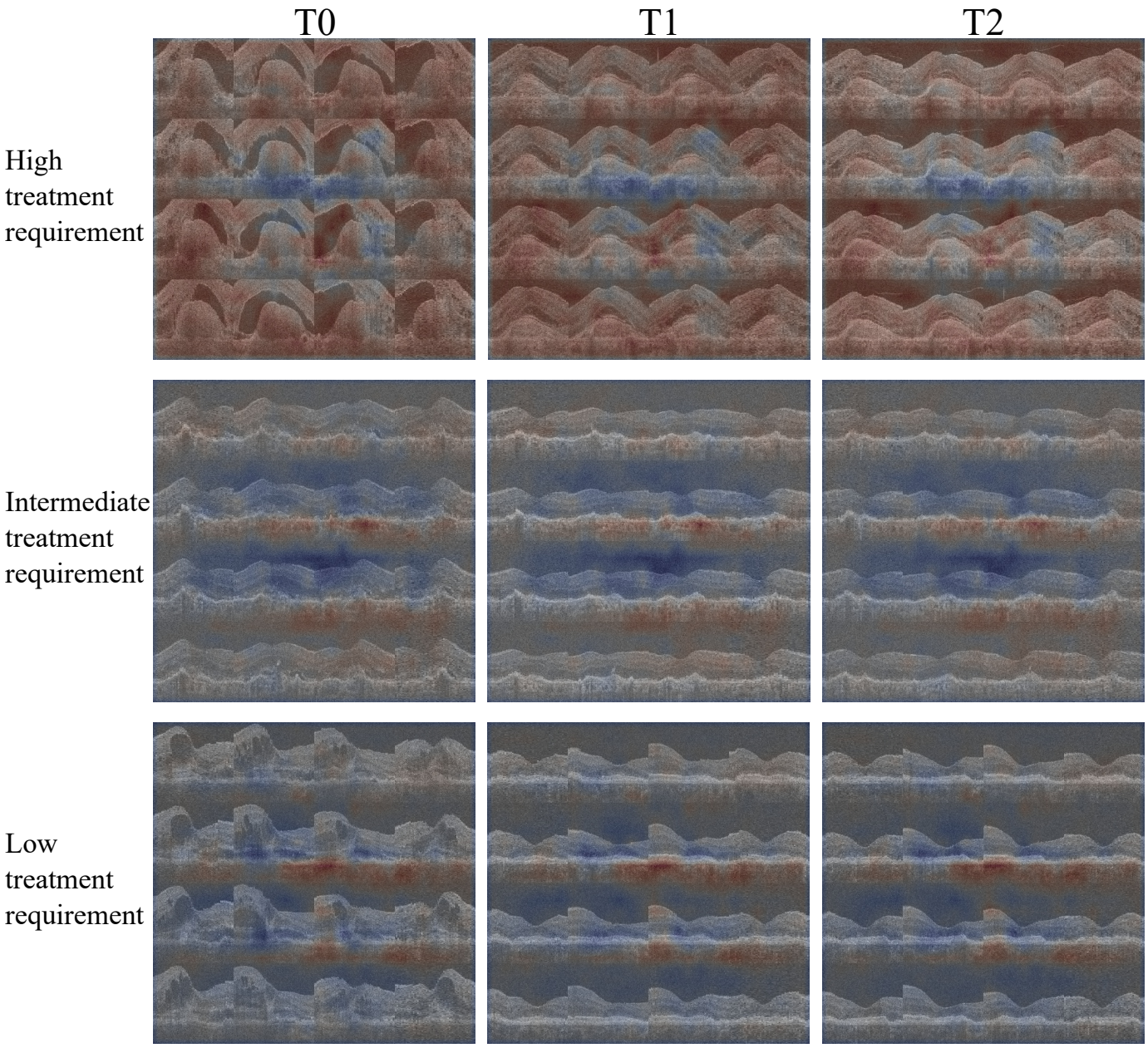

Fig. 5. Example of normalized heatmaps obtained with the occlusion sensitivity visualization technique. For each retina, the same heatmap is superimposed on the multi-tile images associated to the initiation stage. Observe that the retina with high treatment requirement (top row) is strongly deformed in $T0$, and with presence of retinal fluid (black regions). Even in $T2$ some small regions with retinal fluid are still visible. On the other hand, while the low (botton row) and intermediate (center row) also present some fluid in the initial time point, the retinal deformation is smaller in $T0$ and almost imperceptible in $T1$ and $T2$. In the colormap, red region represents those portion of the image where the attribution method assign higher relevance to the decision, while blue regions denote those regions that had lower relevance to the decision.

groups most of the information is gathered from the entire image grid. This is consistent with the fact that, for intermediate-high retreatment retinas, large retinal distortions due to retinal fluid are observed in the image grid even at the last visit of the initiation stage. On the other hand, for the low group the "representative" attribution map points to a specific region of the image grid that lies underneath Bruch's membrane. Differentiation between low-intermediate treatment requirement patients in this case is no longer possible with conventional measures of disease activity (such as retinal fluid), and it is conceivable that the model uses information from that region to support its decision. One hypothesis is that the model was able to capture the transition from angiogenesis to fibrosis, the so called *angiofibrotic switch* [33], in this case occurring under the Bruch's membrane.

Inhibition of VEGF by intravitreal delivery of antibodies was the first treatment to achieve stabilization and/or improvement of visual acuity by resolution of fluid and forged a paradigm-shift in the management of neovascular AMD [34]. However, the functional benefit of anti-VEGF therapy relies strongly on the substances' ability to resolve intra- and subretinal fluid and an efficient long-term monitoring to keep

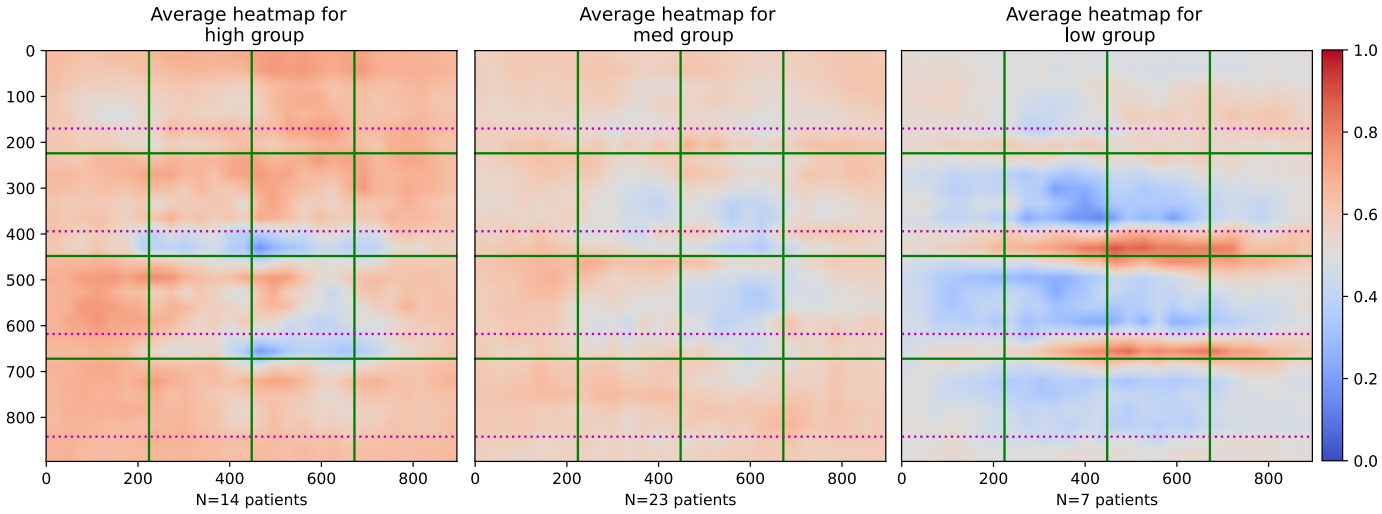

Fig. 6. Average of the normalized heatmaps for the high, intermediate (med) and low treatment requirement groups obtained with the occlusion sensitivity technique. The range for the resulting heatmap was $[0.16 - 0.78]$ for the high, $[0.35 - 0.7]$ for the intermediate, and $[0.14 - 0.9]$ for the low treatment requirement groups. The green grid lines represent the image border for each image tile comprising the 4×4 multi-tile image. The magenta dotted lines represent the level at which Bruch's membrane (BM) was flattened.

the individual retina free of fluid. Substantial discrepancies between individual physicians and certified reading center evaluations resulted in a substantial number of missed treatments in clinical trials using manual detection of recurrent or persistent fluid [35], [21]. The dilemma of OCT misinterpretation in the real world, including the inability to reliably identify, localize and quantify pathological fluid in OCT scans, is associated with an excessive variability in intravitreal injection rates, reimbursement expenses and inferior clinical outcomes [10]. This makes estimating the actual treatment needs difficult and is the reason why we relied on a curated dataset in order to get a more realistic representation of the treatment needs as opposed to solely relying on the number of injections received during a trial.

AI-based methods of automated identification, localization and quantification particularly using end-to-end learning on large OCT volumes in the range of 60-100 million voxels per image are particularly amenable to identify patterns of fluid recurrence [18]. OCT imaging and image analyses can then act as a "VEGF meter" not only detecting fluid, but with predictive tools also identify characteristic patterns of individual disease recurrence as early as after the initiation of therapy in each individual patient. With millions of nAMD patients under life-long therapy world-wide and extensive socioeconomic costs for healthcare physicians' ability to provide individual point-of-care in an AI-controlled setting has a solid potential to introduce a break-through in one of the most frequent therapeutic applications in medicine.

Future work in this topic will include working with larger datasets and finding additional strategies to effectively summarize the high dimensionality of the OCT volumes. The end-to-end DL techniques combined with attribution visualization techniques may help discover novel prognostic imaging patterns as promising imaging biomarkers that would need further clinical investigation.

ACKNOWLEDGMENTS

The financial support by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged. We also thank the NVIDIA corporation for GPU donation.

REFERENCES

[1] N. Ferrara, R. D. Mass, C. Campa, and R. Kim, "Targeting VEGF-A to treat cancer and age-related macular degeneration," *Annu. Rev. Med.*, vol. 58, pp. 491–504, 2007.

[2] M. G. Maguire, D. F. Martin, G.-s. Ying *et al.*, "Five-year outcomes with anti–vascular endothelial growth factor treatment of neovascular age-related macular degeneration: the comparison of age-related macular degeneration treatments trials," *Ophthalmology*, vol. 123, no. 8, pp. 1751–1761, 2016.

[3] R. D. Jager, W. F. Mieler, and J. W. Miller, "Age-related macular degeneration," *NEW ENGL J MED*, vol. 358, no. 24, pp. 2606–2617, 2008.

[4] J. Fujimoto and E. Swanson, "The development, commercialization, and impact of optical coherence tomography," *INVEST OPHTH VIS SCI*, vol. 57, no. 9, pp. OCT1–OCT13, 2016.

[5] M. A. Windsor, S. J. Sun, K. D. Frick, E. A. Swanson, P. J. Rosenfeld, and D. Huang, "Estimating public and patient savings from basic research - a study of optical coherence tomography in managing antiangiogenic therapy," *American Journal of Ophthalmology*, vol. 185, pp. 115–122, jan 2018.

[6] P. J. Rosenfeld, "Optical coherence tomography and the development of antiangiogenic therapies in neovascular age-related macular degeneration," *Investigative Ophthalmology and Visual Science*, vol. 57, no. 9, pp. OCT14–OCT26, jul 2016.

[7] B. G. Busbee, A. C. Ho, D. M. Brown *et al.*, "Twelve-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration," *Ophthalmology*, vol. 120, no. 5, pp. 1046 – 1056, 2013.

[8] M. Okada, R. Kandasamy, E. W. Chong *et al.*, "The treat-and-extend injection regimen versus alternate dosing strategies in age-related macular degeneration: a systematic review and meta-analysis," *American journal of ophthalmology*, vol. 192, pp. 184–197, 2018.

[9] T. A. Ciulla, F. Huang, K. Westby *et al.*, "Real-world outcomes of anti-vascular endothelial growth factor therapy in neovascular age-related macular degeneration in the United States," *Ophthalmology Retina*, vol. 2, no. 7, pp. 645–653, Jul. 2018.

[10] H. Mehta, A. Tufail, V. Daien *et al.*, "Real-world outcomes in patients with neovascular age-related macular degeneration treated with intravitreal vascular endothelial growth factor inhibitors," *PROG RETIN EYE RES*, vol. 65, pp. 127–146, 2018.

[11] D. S. Kermany, M. Goldbaum, W. Cai *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[12] J. De Fauw, J. R. Ledsam, B. Romera-Paredes *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *NAT MED*, vol. 24, no. 9, p. 1342, 2018.

[13] R. Rasti, M. J. Allingham, P. S. Mettu, S. Kavusi, K. Govind, S. W. Cousins, and S. Farsiu, "Deep learning-based single-shot prediction of differential effects of anti-VEGF treatment in patients with diabetic macular edema," *Biomedical Optics Express*, vol. 11, no. 2, p. 1139, Feb. 2020.

[14] D. B. Russakoff, A. Lamin, J. D. Oakley, A. M. Dubis, and S. Sivaprasad, "Deep learning for prediction of AMD progression: A pilot study," *Investigative Ophthalmology and Visual Science*, vol. 60, no. 2, pp. 712–722, feb 2019. [Online]. Available: http://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.18-25325

[15] J. Zhao, Q. Feng, P. Wu *et al.*, "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction," *Scientific reports*, vol. 9, no. 1, p. 717, 2019.

[16] T. Wang, R. G. Qiu, and M. Yu, "Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks," *Scientific reports*, vol. 8, 2018.

[17] R. Cui and M. Liu, "RNN-based longitudinal analysis for diagnosis of Alzheimer's disease," *COMPUT MED IMAG GRAP*, vol. 73, pp. 1 – 10, 2019.

[18] U. Schmidt-Erfurth, H. Bogunovic, A. Sadeghipour *et al.*, "Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration," *Ophthalmology Retina*, vol. 2, no. 1, pp. 24–30, 2018.

[19] W.-D. Vogl, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and G. Langs, "Predicting Macular Edema Recurrence from Spatio-Temporal Signatures in Optical Coherence Tomography Images." *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1773–1783, sep 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7915767/ http://www.ncbi.nlm.nih.gov/pubmed/28475051

[20] H. Bogunovic, S. M. Waldstein, T. Schlegl *et al.*, "Prediction of anti-VEGF treatment requirements in neovascular AMD using a machine learning approach," *INVEST OPHTH VIS SCI*, vol. 58, no. 7, p. 3240, 2017.

[21] C. A. Toth, F. C. DeCroos, G. J. Jaffe *et al.*, "Comparison of RC and investigator determined retinal and subretinal fluid in the comparison of age-related macular degeneration treatment trials (CATT)," *INVEST OPHTH VIS SCI*, vol. 53, no. 14, pp. 2894–2894, 2012.

[22] T. Schlegl, S. M. Waldstein, H. Bogunovic *et al.*, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, Apr. 2018.

[23] A. Montuoro, J. Wu, S. Waldstein *et al.*, "Motion artefact correction in retinal optical coherence tomography using local symmetry," in *MICCAI*, ser. LNCS, vol. 8674, 2014, pp. 130–137.

[24] M. K. Garvin, M. D. Abramoff, X. Wu *et al.*, "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE T MED IMAGING*, vol. 28, no. 9, pp. 1436–1447, 2009.

[25] K. Li, X. Wu, D. Z. Chen, and M. Sonka, "Optimal surface segmentation in volumetric images-a graph-theoretic approach," *IEEE T PATTERN ANAL*, vol. 28, no. 1, pp. 119–134, 2006.

[26] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. of the IEEE CVPR conference*, vol. 1, no. 2, 2017, p. 3.

[27] H. Li, Z. Xu, G. Taylor *et al.*, "Visualizing the loss landscape of neural nets," in *Proc. of the NIPS conference*, 2018, pp. 6391–6401.

[28] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[30] P. Haeberli and D. Voorhies, "Image processing by linear interpolation and extrapolation," *IRIS Universe Magazine*, vol. 28, pp. 8–9, 1994.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[33] P. K. Roberts, S. Zotter, A. Montuoro *et al.*, "Identification and quantification of the angiofibrotic switch in neovascular AMD," *INVEST OPHTH VIS SCI*, vol. 60, no. 1, pp. 304–311, 01 2019.

[34] P. J. Rosenfeld, D. M. Brown, J. S. Heier *et al.*, "Ranibizumab for neovascular age-related macular degeneration," *NEW ENGL J MED*, vol. 355, no. 14, pp. 1419–1431, 2006.

[35] D. F. Martin, M. G. Maguire, S. L. Fine, G. shuang Ying, G. J. Jaffe, J. E. Grunwald, C. Toth, M. Redford, and F. L. Ferris, "Ranibizumab and bevacizumab for treatment of neovascular age-related macular degeneration: Two-year results," *Ophthalmology*, vol. 119, no. 7, pp. 1388 – 1398, 2012.