# Prediction of treatment response in patients with neovascular age-related macular degeneration

Anil Rao[1,2], Shruti Chandra[1,2], and Sobha Sivaprasad[1,2]

[1] NIHR Moorfields Biomedical Research Centre,
Moorfields Eye Hospital, London, United Kingdom
[2] UCL Institute of Ophthalmology,
11-43 Bath Street, London EC1V 9EL, United Kingdom

**Abstract.** Neovascular age-related macular degeneration (nAMD) is a common cause of visual impairment, and is currently treated with intravitreal anti-vascular endothelial growth factor agents such as aflibercept. While these treatments may improve visual acuity (VA) in some patients, clinicians cannot currently predict who is likely to benefit before treatment starts. The aim of this study is to explore the effectiveness of using Deep Learning approaches to train models for predicting whether a patient's VA will respond favourably to three months of aflibercept therapy, using pre-treatment OCT images and clinical/demographic variables. We train a number of models using standard machine learning, Deep Learning transfer learning, and fully trained Deep Learning approaches in two experiments using outcomes based on the VA at 4-10 weeks after the final dose. In experiment one, we trained models to predict whether the VA will be at least 54 Early Treatment Diabetic Retinopathy Study (ETDRS) letters, while in experiment two we trained them to predict whether the VA will have increased by 10 or more letters. Model prediction quality was assessed using the Area Under the Curve (AUC) of the Receiver-Operating-Characteristic (ROC) curves. We found that all models performed significantly better than chance in both experiments, except for the fully trained Deep Learning model using just images in experiment two. The best performing model for experiment one was the Deep Learning transfer model using images and clinical/demographic variables (AUC=0.901), while in experiment two, none of the Deep Learning approaches performed better than a random forest using only clinical/demographic variables (AUC=0.751). Our experiments suggest that different Deep Learning approaches are required for predicting the second outcome if we want the models to perform better than those that use clinical/demographic variables alone.

**Keywords:** Neovascular age-related macular degeneration (nAMD) · Optical Coherence Tomography (OCT) · Deep Learning · Anti-VEGF · Aflibercept · Treatment Response.

## 1 Introduction

Neovascular age related macular degeneration (nAMD) remains a common cause of visual impairment amongst older individuals. The condition is characterised by the onset of macular neovascularisation that can leak fluid and/or blood and cause distortion of the macular architecture and consequent visual impairment of varying degrees. It is typically first diagnosed by an optometrist using Optical Coherence Tomography (OCT) of the eye, before confirmation of the diagnosis in hospital using OCT/fluorescein angiography.

Currently, nAMD is treated using intravitreal anti-vascular endothelial growth factor agents, such as aflibercept. While such agents aim to resolve the macular fluid, this does not necessarily correspond to an improvement in the visual acuity of treated subjects. The standard treatment regimen of aflibercept is to load with three months of intravitreal injection, before deciding on further treatment based on how the patient's macular anatomy and visual acuity have responded. It would be useful to automatically predict the visual acuity response based on the patients initial presentation, not only to manage patient expectations but also because it is a predictor of patient outcomes after 12 months. However, it remains challenging to perform such a prediction in routine clinical practice.

As a result, a number of investigators have attempted to identify or describe possible biomarkers for treatment response in patients with nAMD [5, 13, 9, 6, 18]. Some of these biomarkers such as intraretinal cystoid fluid, subretinal fluid and hyperreflective foci, are present on OCT and may be associated with visual acuity before and after treatment with intravitreal therapy [18]. However, robust and accurate segmentation algorithms are required to extract these biomarkers from OCT if we wish to use them as quantitative features for the prediction of treatment response.

An alternative to using biomarkers for predictive modelling is to use Deep Learning methods, whereby image features are extracted by training a Convolutional Neural Network (CNN). Deep Learning has been used in a number of prediction tasks involving OCT, such as disease classification [14, 3], prediction of conversion to nAMD [24], and treatment referral [16, 11]. In [23], Deep Learning was used to predict treatment response of patients with nAMD from OCT and clinical/demographic variables. Although the trained models achieved high predictive accuracy, the study design differed from that of this work, and the described approach used a patented Neural Network architecture which therefore has a restricted use.

The aim of this work is to use Deep Learning approaches to try to predict treatment response of eyes with treatment-naive nAMD using routinely obtained OCT and clinical/demographic variables acquired just prior to treatment. In the first approach, transfer learning is used to extract image features from OCT, which are then combined with the clinical/demographic variables to give the input features for training machine learning algorithms. In the second approach, fully trained Deep Learning models are developed using only the OCT images. In contrast with [23], we use outcome measures based on visual acuity/changes in visual acuity 4-10 weeks after the last loading phase injection. The models are developed on a random training sample of data from a multi-site study, and evaluated on the remaining unseen test sample.

Section 2 now describes the study data used for analysis before we describe the methods used for preprocessing and machine learning in section 3. This is followed by a presentation of the results in section 4 before a concluding discussion in section 5.

## 2 Data

The PRECISE study is a multi-site study that looks at imaging markers to predict treatment response to the loading phase of intravitreal aflibercept therapy, in subjects with treatment naive nAMD. Please see section 5 for the list of study collaborators. The study design consists of retrospective and prospective data from four visits, with mandatory Heidelberg OCT obtained at visits one and four. (Heidelberg OCTA is also optionally obtained but is not used in this analysis.) All patients undergo a total of three monthly loading injections of intravitreal aflibercept therapy over visits one to three, and visit four occurs 4-10 weeks after the third loading injection.

The inclusion criterion for the study were: (1) Adults between 50 and 100 years, (2) Treatment naive nAMD at baseline, (3) Media clarity, pupillary dilation and patient cooperation for adequate imaging. (4) Ability to give informed consent. The retrospective part of the study had additional inclusion criteria which essentially ensured that the visit schedule and available data matched the study design. Note that more than one eye from a patient could be entered into the study and they were considered separate subjects. The exclusion criteria were: (1) Co-existent ocular disease: Any other ocular condition that, in the opinion of the investigator, might affect or alter visual acuity during the course of the study. (2) Any patient who has opted out of their information being used for research nationally or locally at any site.

A set of 2000 OCT eye volumes from visit one were exported for analysis from this study. These volumes were from 1865 unique patients. As these volumes were not always centred on the fovea, a clinical expert marked the fovea using Heyex software before export, so that it could be determined from the image metadata. In addition, the following clinical/demographic variables were available: Visual Acuity (VA) at visits one and four ($\mathrm{VA}^1$ and $\mathrm{VA}^4$), Age, Gender (Male/Female), and Ethnicity (White/Black/South Asian/Other Asian/Other). Visual Acuity was recorded as the number of Early Treatment Diabetic Retinopathy Study (ETDRS) letters and lies in the range 0–100.

## 3 Methods

### 3.1 Image Preprocessing

As the exported OCT volumes were heterogeneous in terms of resolution and eye coverage, they were standardized to a uniform resolution and coverage before further analysis. In what follows, the OCT axes are defined as x: patient right to patient left, y: posterior to anterior and z: inferior to superior. Firstly, the foveal x- and z-coordinates within each volume were extracted from the volume metadata using custom-written software. Then, each b-scan was independently downsampled using linear interpolation to a resolution of 0.02 mm × 0.008 mm and laterally cropped to a size of 240 × 240. The lateral cropping was performed symmetrically about the column index of the a-scan corresponding to the foveal x-coordinate. The volumes were then downsampled using nearest-neighbour interpolation in the z-direction to a resolution of 0.5 mm, and cropped giving a new volume consisting of seven b-scans. Nearest neighbour interpolation was used in the z-direction to prevent smoothing away of image detail due to the relatively large distance between b-scans [24]. The cropping in the z-direction was performed symmetrically about the foveal z-coordinate of the OCT volume. Note that due to heterogeneity in the foveal locations and coverage of the original OCT volumes, 63 volumes could not be cropped to the standardized coverage, and so were not included in the analysis. In addition, three volumes that had been acquired at an angle of greater than 10 degrees tilt to the z-axis were excluded, in order to ensure approximately uniform orientation of the standardized volumes. One further volume was excluded due to inconsistencies in the distances between b-scans at time of acquisition.

The resulting standardized and cropped volumes of any left eyes were then laterally flipped to minimize variation in the data. Finally, data was discarded in which the image quality of the b-scan containing the fovea was less than 15. The cutpoint of 15 was chosen because it corresponds to a threshold for 'medium-quality' images [2].

### 3.2 Outcome Variables

Two outcome variables were used to measure the treatment response of each subject to intravitreal aflibercept therapy. The first outcome variable, $y^v$ is based on the VA at visit four. It is defined as follows:

$$y_i^v = \begin{cases} -1: & \text{if } \mathrm{VA}_\mathrm{i}^4 < 54 \\ 1: & \text{if } \mathrm{VA}_\mathrm{i}^4 \geq 54 \end{cases} \tag{1}$$

In the above, $y_i^v = -1$ is considered a treatment non-responder, and $y_i^v = 1$ is considered a treatment responder. (In what follows, we will use the abbreviations 'responder', and 'non-responder' for clarity.) The cutpoint of 54 letters was chosen because it separates those with severe/moderate visual impairment from those that have mild visual impairment/driving vision.

While this provides a measure of the visual function at visit four, it does not *directly* quantify treatment response because different subjects have different VA at visit one, i.e., before treatment is started. To give a more direct measure of treatment response, we use the following outcome variable, $y^{\delta v}$ based on the change in VA between the first and fourth visits:

$$y_i^{\delta v} = \begin{cases} -1: & \text{if } \mathrm{VA}_\mathrm{i}^4 - \mathrm{VA}_\mathrm{i}^1 < 10 \\ 1: & \text{if } \mathrm{VA}_\mathrm{i}^4 - \mathrm{VA}_\mathrm{i}^1 \geq 10 \end{cases} \tag{2}$$

As before, $y_i^{\delta v} = -1$ is considered a non-responder, and $y_i^{\delta v} = 1$ is considered a responder. An increase in VA of 10 letters was chosen as the cutpoint because it is considered to represent a clinically meaningful improvement in subjects with advanced eye disease [12]. This corresponds to an increase in VA of two lines of the Snellen chart, which is the cutpoint used in [23] for prediction of treatment response from OCT images and clinical/demographic variables.

### 3.3 Clinical/Demographic Variables Models

We firstly train models using just the clinical/demographic variables of each subject for prediction of both outcomes. It is important to evaluate these models because variables such as $\mathrm{VA}^1$ and Age have been previously shown to be associated with treatment-related gains in VA [18].

**Logistic Regression (C)** We build a logistic regression model using only clinical/demographic features as the model inputs. The clinical/demographic features vector $c_i$ consists of $\mathrm{VA}^1$, Age, Gender (Male/Female), and Ethnicity (collapsed to Non-White/White). The categorical variables Gender and Ethnicity are both coded as 0/1, with Male=0, Female=1 and Non-White=0, White=1, respectively. The model parameters $\beta, \beta_0$ are then the minimizers of the weighted logistic loss:

$$\sum_{i=1}^{n} w_i \log \left[ e^{-y_i(\widehat{c_i}\beta + \beta_0)} + 1 \right] \tag{3}$$

in which $n$ is the size of the training set and $\widehat{c_i}$ are the clinical/demographic features vector $c_i$ standardized across subjects to zero mean and unit variance. The target variable $y$ is either $y^v$ or $y^{\delta v}$, depending on the prediction task. The weights $w_i$ are used to weight the loss of each example and are defined as

$$w_i = \frac{n}{2n_i} \tag{4}$$

where $n_i$ is the number of training examples from class $y_i$. Such a weighting is often used to mitigate the effect of class-imbalance on algorithm training.

**Random Forest ($\mathbf{C^r}$)** Here we build a model using only $c_i$ as the model inputs to a Random Forest classifier [21]. These are ensemble classifiers in which a number of decision trees are built using bootstrap samples from the training data. At each node of the tree, a random subset of input features are selected and the node and splitting threshold which maximises the improvement in Gini Impurity [4] is chosen. For our purposes, the random forest is trained using class-balancing sample weights as in equation 4. Random forests have a number of hyperparameters such as the number of trees, the maximum tree depth, $max\_depth$, and the number of features to be considered at each split, $max\_features$.

### 3.4 Transfer Learning Models

The first Deep Learning approach we use in this study is transfer learning. Here, we utilize the VGG19 Convolutional Neural Network (CNN) model [20] available as part of the Keras API within TensorFlow (v2.0) [1]. This model has been shown to achieve competitive image classification performance when trained on the ImageNet database [7], which is a collection of millions of images with labels such as cat, dog, aeroplane etc. Since Keras provides the VGG19 model and its model weights after pre-training on ImageNet, this facilitates the use of transfer learning to develop a new classifier specific to our task.

Figure 1 shows how transfer learning is operationalized for treatment response prediction. Firstly, the VGG19 model, with pre-trained ImageNet weights, is downloaded using Keras with the option 'include_top = False'. This option ensures that the downloaded model does not include the Dense and Softmax layers at the head of the original model which are specific to classifying the ImageNet data. Instead, we add a Global Average Pooling layer which takes the 512 features from the final block and averages them over the image. We also incorporate preprocessing layers to prepare the image inputs for the VGG19 model. Finally, all layers in the resulting base model are 'frozen' as non-trainable layers.

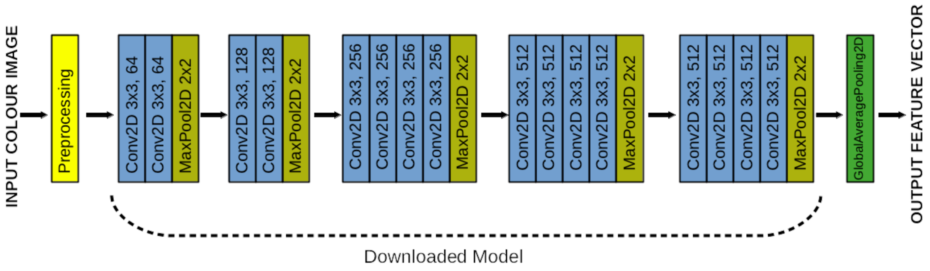


Fig. 1: This figure shows how the base VGG19 model is constructed using Keras. The base model is then used as a feature extractor for each b-scan.

The resulting base model can then be used as a feature extractor to produce feature vectors of length 512 for any 2D colour image. As each b-scan consists of a single channel, we simply replicate the b-scan 3 times to give a pseudo-colour image that can be input into the base model. Application of the feature extractor to every b-scan from the $i$th subject gives the following feature vector $x_i$:

$$x_i \equiv [x_i^{(1)}, \ldots, x_i^{(7)}] \tag{5}$$

Here, $x_i^{(j)}$ is the concatenation of the base model outputs $x_i^{(j)}$ of the corresponding b-scans $B_i^{(j)}$. Each feature vector $x_i$ is therefore a vector of length 3584 $(= 512 \times 7)$ which represents the OCT volume for subject $i$. These features can then be used together with the clinical/demographic variables as inputs to standard machine learning algorithms for learning the relationship between the OCT volume and/or the clinical/demographic variables, and treatment response. For our purposes we use the following algorithms for predicting each of the outcome variables.

**Images Only: $l_2$-Regularized Logistic Regression (I)** In this model, we use the image features $x_i$ from equation (5) as the model inputs. The model parameters $\beta, \beta_0$ are the minimizers of the penalized weighted logistic loss:

$$\frac{||\beta||_2^2}{2} + C\sum_{i=1}^{n} w_i \log\left[e^{-y_i(\widehat{x_i}\beta+\beta_0)} + 1\right] \tag{6}$$

where $||\cdot||_2$ is the $l_2$ norm, $\widehat{x_i}$ are $x_i$ standardized across subjects to zero mean and unit variance, and C is a hyperparameter which is inversely proportional to the regularization strength.

**Images Only: $l_2$-Regularized Logistic Regression with kernel smoothing ($\mathbf{I^k}$)** In this model, the model inputs are kernel-smoothed versions of the image features $x_i$. The model parameters $\beta, \beta_0$ are the minimizers of the weighted logistic loss:

$$\frac{||\beta||_2^2}{2} + C\sum_{i=1}^{n} w_i \log\left[e^{-y_i(\widehat{x_i^S}\beta+\beta_0)} + 1\right] \tag{7}$$

where $\widehat{x_i^S}$ are standardized versions of $x_i^S$, which are the OCT image features at the foveal b-scan, $x_i^{(4)}$, after kernel-smoothing in the z-direction:

$$x_i^S = \sum_{j=1}^{7} e^{-\frac{(j-4)^2}{2s^2}} x_i^{(j)} \tag{8}$$

In the above, $s$ is an additional hyperparameter that controls the degree of smoothing: Small values of $s$ correspond to less smoothing in which $x_i^S \approx x_i^{(4)}$, while large values of $s$ correspond to greater smoothing in which $x_i^S \approx \sum_{j=1}^{7} x_i^{(j)}$. Note that since we use the standardized features $\widehat{x_i^S}$ in equation (7), there is no explicit normalisation in equation (8).

The motivation for the above approach derives from the fact that the image features $x_i^{(j)}$ are 2D features that potentially capture whether a high-level abstraction, such as fluid, occurs in the corresponding b-scan $B_i^{(j)}$. Equation (8) essentially pools this information by forming a weighted average of those features over the entire OCT volume. Moreover, the weighting scheme encodes the prior knowledge that the foveal features $x_i^{(4)}$ are the most relevant to the classification task, and so are given the greatest weight, while features for b-scans relatively far from the foveal b-scan are less relevant, and so are given smaller weights.

**Images & Clinical/Demographic Variables: $l_2$-Regularized Logistic Regression with feature scaling ($\mathbf{IC^f}$)** In this model, we utilize both the image features $x_i$ and the clinical/demographic features $c_i$ as model inputs. The model parameters $\beta, \beta_0$ are the minimizers of the penalized weighted logistic loss:

$$\frac{||\beta||_2^2}{2} + C\sum_{i=1}^{n} w_i \log\left[e^{-y_i([\widehat{x_i},\lambda\widehat{c_i}]\beta+\beta_0)} + 1\right] \tag{9}$$

where $[\widehat{x_i}, \lambda\widehat{c_i}]$ is a concatenation of the standardised image feature vector $\widehat{x_i}$ and the standardised clinical/demographic features vector $\widehat{c_i}$ multiplied by a hyperparameter $\lambda > 0$. Due to the regularisation term $\frac{||\beta||_2^2}{2}$, $\lambda$ effectively controls the relative importances of the image and clinical/demographics feature vectors in the objective function. This can be considered as a kind of Multiple Kernel Learning [10], in which the kernel weights are hyperparameters rather than parameters that are optimised during model training.

**Images & Clinical/Demographic Variables: $l_2$-Regularized Logistic Regression with kernel smoothing and feature scaling ($\mathbf{IC^{kf}}$)** The final transfer learning model utilizes kernel-smoothed versions of the image features $x_i$ and the clinical/demographic features $c_i$ as the model inputs. The model parameters $\beta, \beta_0$ are the minimizers of the penalized weighted logistic loss:

$$\frac{||\beta||_2^2}{2} + C\sum_{i=1}^{n} w_i \log\left[e^{-y_i([\widehat{x_i^S},\lambda\widehat{c_i}]\beta+\beta_0)} + 1\right] \tag{10}$$

### 3.5 Fully Trained Deep Learning Models

In the second Deep Learning approach, we train complete Deep Learning models from scratch. For this purpose, we construct the CNN model shown in figure 2 using Keras. This model takes as inputs the preprocessed b-scans $B_i^{(j)}$ only. In the first step, the image values of the seven preprocessed b-scans of a subject are divided by 255 so that they are in the range 0–1. Each rescaled scan subsequently passes through two identical blocks consisting of thirty-two 2D Convolutional filters of size 3×3, a Batch Normalization layer and a MaxPooling layer of size 2×2. They then pass through an additional block with similar Convolutional/Batch Normalization layers but with a MaxPooling layer of size 4×4. A Global Average Pooling layer is then applied followed by a Dense layer consisting of 32 units. At this point, the seven outputs from the Dense layer are averaged using an Average layer and passed through a Dropout layer with dropout rate of 0.5, and a Dense layer with a single unit and sigmoid activation. 'Relu' activation and $l_2$ weight-regularization is used in each of the convolutional layers. We refer to this model as $\mathrm{I^d}$.

The CNN architecture shown in figure 2 differs to the typical Deep Learning architectures used in Ophthalmology for prediction. Usually, 2D CNNs are used in which the model input is a single b-scan from a subject [3, 23, 16, 17]. Such approaches require a way of pooling the model predictions from each b-scan to give a volume-level prediction. Alternatively, 3D CNNs are used in which an entire OCT volume (or sub-volume) is the model input [24, 11]. In our approach we have used a 2D CNN in which there are seven model inputs, $B_i^{(j)}$, for each subject rather than one. Model training up to the Average layer amounts to estimating a single set of weights that is applied to each of the model inputs independently, until the Average layer 'pools' the information across inputs by taking the mean. The motivation is similar to that for model $\mathrm{I^k}$: We assume that the layers preceding the Average layer will form a high-level abstraction of each b-scan that can then be averaged across scans, and used in the following layers for fitting to the outcome of interest. Our approach, unlike a typical 2D CNN, therefore uses all the scans of a subject simultaneously during the fitting process to directly give volume-level model predictions. This is preferable because, while the signal that distinguishes between a non-responder and a responder may be present in an OCT volume, it may not be present in all of its constituent b-scans. Furthermore, our approach does not contain more parameters than a 2D CNN, unlike approaches that use 3D CNNs. This is advantageous where the training set size is relatively small, as in our case.

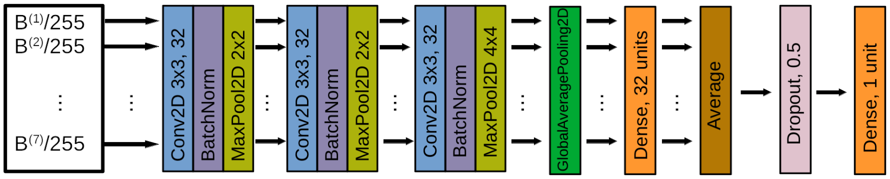


Fig. 2: This figure shows how the fully trained Deep Learning CNN is constructed using Keras. The model takes seven inputs which are the rescaled versions of the seven preprocessed b-scans of a subject. Note that, up to the Average layer, a *single* set of weights is estimated and independently applied to each of the seven inputs.

### 3.6 Evaluation Methodology

The set of 2000 OCT volumes was randomly divided into a training set of size 1333, and a test set of size 667. This was performed while ensuring that volumes from the same patient appeared in only one of the training or test sets. At the time of analysis, 1992 of these 2000 baseline OCT volumes were considered eligible for analysis after clinical grading was performed. The final training and test sets for prediction of $y^v$ were then given by additionally excluding volumes for reasons described in section 3.1, and for missing clinical/demographic and outcome variables. This results in a training set of size 1228, from 1142 unique patients, and a test set of size 619, from 580 unique patients, for prediction of $y^v$. For prediction of $y^{\delta v}$, we further discard volumes in which $\text{VA}^1 \geq 91$, since it is not possible for those subjects to increase their VA by 10 letters as $\text{VA} \leq 100$. This gives a training set of size 1226, from 1140 unique patients, and a test set of size 619, from 580 unique patients, for prediction of $y^{\delta v}$. Table 1 gives the distribution of clinical/demographic variables and outcomes in the training and test sets used for prediction of each outcome variable. Note that all model training and evaluation was performed at the subject level for both outcomes.

(a) $y^v$

| **Variable** | | **Training** | **Test** |
|---|---|---|---|
| Number of Subjects | | 1228 | 619 |
| Number of Unique Patients | | 1142 | 580 |
| $\text{VA}^1$ | | $57.9 \pm 14.9$ | $58.6 \pm 14.3$ |
| Age | | $79.4 \pm 7.7$ | $79.4 \pm 7.8$ |
| Gender: Male/Female | (%) | 40.1/59.9 | 37.2/62.8 |
| Ethnicity: Non-White/White | (%) | 4.3/95.7 | 5.0/95.0 |
| $\text{VA}^4$ | | $62.4 \pm 14.9$ | $62.8 \pm 15.4$ |
| Outcome: Non-Responder/Responder | (%) | 24.3/75.7 | 21.3/78.7 |

(b) $y^{\delta v}$

| **Variable** | | **Training** | **Test** |
|---|---|---|---|
| Number of Subjects | | 1226 | 619 |
| Number of Unique Patients | | 1140 | 580 |
| $\text{VA}^1$ | | $57.9 \pm 14.9$ | $58.6 \pm 14.3$ |
| Age | | $79.4 \pm 7.7$ | $79.4 \pm 7.8$ |
| Gender: Male/Female | (%) | 40.1/59.9 | 37.2/62.8 |
| Ethnicity: Non-White/White | (%) | 4.3/95.7 | 5.0/95.0 |
| $\text{VA}^4$ | | $62.4 \pm 14.9$ | $62.8 \pm 15.4$ |
| Outcome: Non-Responder/Responder | (%) | 70.5/29.5 | 73.0/27.0 |

Table 1: This table shows the distribution of clinical/demographic variables and outcomes in the training and test sets used for prediction of $y^v$ and $y^{\delta v}$. For continuous variables, the mean and standard deviation are given. Note that all statistics are calculated at the subject, rather than patient, level.

Each of the clinical/demographic variables and transfer learning models was trained using the features and outcome from the training set. Standardization of features was performed using only the training data and then applied to the training and test data. The optimum hyperparameters for the models I, $\text{I}^{\text{k}}$, $\text{IC}^{\text{f}}$ and $\text{IC}^{\text{kf}}$ were chosen using a grid-search as those that minimize the weighted logistic loss in 10-fold cross validation of the training set. The range of hyperparameters in the grid were: $C = 10^r$, $r \in [-5, -4, \ldots, 5]$; $\lambda = 2^r$, $r \in [-6, -5, \ldots, 8]$; $s = \frac{3}{\sqrt{2}} \times 1.5^r$, $r \in [-5, -4, \ldots, 5]$. The model estimation and grid-search for each of these models was performed using scikit-learn (v0.23.2) [15]. Model estimation for model $C$, which contains no hyperparameters, was performed using the statsmodels package (v0.12.2) [19]. Once all parameters and hyperparameters have been estimated using the training data, a test example with image features $x_t$ and clinical features $c_t$ is given the following predicted probability of being a responder:

$$P(y_t = 1) = \frac{1}{1 + e^{-f_t \beta + \beta_0}} \tag{11}$$

where $f_t$ equals $\widehat{c_t}$ for model C, $\widehat{x_t}$ for model I, $\widehat{x_t^S}$ for model $\mathrm{I^k}$, $[\widehat{x_t}, \lambda\widehat{c_t}]$ for model $\mathrm{IC^f}$, and $[\widehat{x_t^S}, \lambda\widehat{c_t}]$ for model $\mathrm{IC^{kf}}$. The test example is given a binary prediction of responder if $P(y_t = 1) > 0.5$, and non-responder otherwise. For model $\mathrm{C^r}$, the number of trees was fixed to 1000, and optimum values for *max_depth* and *max_features* were chosen using a grid-search as those that maximize the balanced accuracy in 10-fold cross validation of the training set. The range of hyperparameters in the grid were: $max\_depth \in [1, 2, \ldots 11]$ and $max\_features \in [2, 3, 4]$. Model estimation and grid-search for $\mathrm{C^r}$ was performed using scikit-learn. After all hyperparameters have been estimated and the random forest has been trained, a test example with clinical features $c_t$ is passed through each decision tree in the forest and a probability of being a responder, $P(y_t = 1)$ is calculated. The test example is given a binary prediction of responder if $P(y_t = 1) > 0.5$, and non-responder otherwise.

The CNN model $\mathrm{I^d}$ was trained on a stratified sample of 80% of the preprocessed images from the training set. This was performed using the Adam optimizer and a weighted logistic loss function with weights given by equation (4). The model was trained for 300 epochs using a learning rate scheme in which the initial learning rate was divided by 10 after 10 epochs and then divided by 10 again after a further 10 epochs. The remaining 20% of the training data was used as a validation set to determine the optimum values for the $l_2$ weight-regularization $\alpha$ and the initial learning rate $lr$ using the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curves as the validation metric. This was achieved using early-stopping and a grid-search over the values $\alpha = 5 \times 10^r$, $r \in [-7, -6, \ldots, -3]$; $lr = 10^r$, $r \in [-5, -4, \ldots, -1]$. Note that the 'patience' for the early-stopping was set to infinite, so this procedure effectively optimizes for the number of training epochs between 1 and 300. A test example is passed through the optimum model to give it's probability of being a responder, $P(y_t = 1)$. Since we have used the AUC to choose the best model, we optimize the probability threshold, $T$, that is used to give a binary prediction of non-responder/responder rather than use the default value of 0.5. We do this by choosing the value that maximizes the geometric mean of sensitivity and specificity on the validation data, with the intention of reducing the disparity between sensitivity and specificity, while maintaining high values for each of them. The test example is then given a binary prediction of responder if $P(y_t = 1) > T$, and non-responder otherwise. Note that in our experiments, the optimization of $T$ did not appear necessary for the prediction of $y^v$, as the evaluated sensitivity and specificity on the validation set were almost identical to each other. Moreover, the optimization of $T$ increased, rather than decreased, the difference between sensitivity and specificity for the prediction of $y^v$ on the validation data. In contrast, the optimized $T$ greatly improved the balance of sensitivity and specificity on the validation data for the prediction of $y^{\delta v}$. Nevertheless, we optimize $T$ for both $y^v$ and $y^{\delta v}$ for consistency and reproducibility of the procedure used for tuning the Deep Learning models used in this work.

Prediction quality was determined using the AUC for each model. In our experiments, the AUC of a model can be interpreted as the probability that it predicts a higher probability of responding to treatment for a randomly chosen responder than a randomly chosen non-responder. In addition, confidence intervals for the AUC are constructed to quantify its uncertainty due to sampling variability. These are determined using an implementation of the percentile bootstrap based on the scipy [22] package.

We also calculate the sensitivity and specificity of each of the models to provide further information about their performance. Sensitivity is defined as the proportion of true responders identified by the model, while specificity is the proportion of true non-responders identified. Clopper-Pearson confidence intervals for sensitivity and specificity are determined using the 'proportion_confint' function available in the statsmodels package.

## 4 Results

### 4.1 Prediction of $y^v$

Table 2: Model performance for prediction of $y^v$

| **Model** | **AUC** | **(95% CI)** | **Sensitivity** | **(95% CI)** | **Specificity** | **(95% CI)** |
|---|---|---|---|---|---|---|
| C | 0.891 | (0.860–0.919) | 0.811 | (0.773–0.845) | 0.826 | (0.750–0.886) |
| $\mathrm{C^r}$ | 0.883 | (0.849–0.914) | 0.821 | (0.784–0.854) | 0.795 | (0.717–0.861) |
| I | 0.787 | (0.741–0.829) | 0.739 | (0.698–0.778) | 0.697 | (0.611–0.774) |
| $\mathrm{I^k}$ | 0.764 | (0.715–0.810) | 0.737 | (0.696–0.776) | 0.659 | (0.572–0.739) |
| $\mathrm{IC^f}$ | 0.901 | (0.871–0.927) | 0.830 | (0.793–0.862) | 0.803 | (0.725–0.867) |
| $\mathrm{IC^{kf}}$ | 0.896 | (0.865–0.924) | 0.828 | (0.791–0.860) | 0.795 | (0.717–0.861) |
| $\mathrm{I^d}$ | 0.772 | (0.727–0.814) | 0.801 | (0.763–0.835) | 0.583 | (0.494–0.668) |

Table 2 gives the performance of each model for prediction of $y^v$, and figure 3 shows the corresponding ROC curves. Every model has an AUC with a 95% confidence interval whose lower bound is greater than 0.5. This indicates that all these models perform better than chance with respect to whether a randomly chosen responder, $y^v = 1$, is given a higher probability of responding to treatment

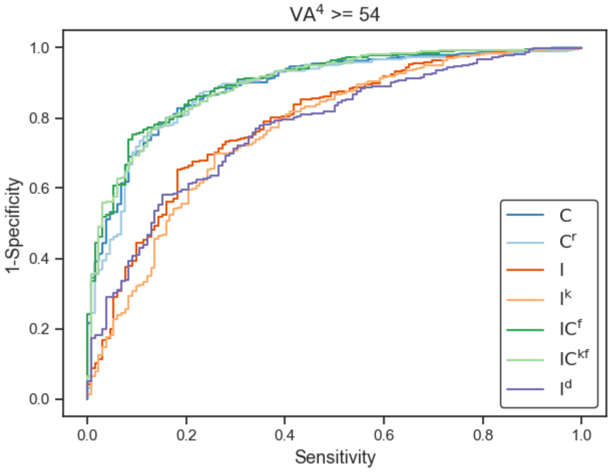
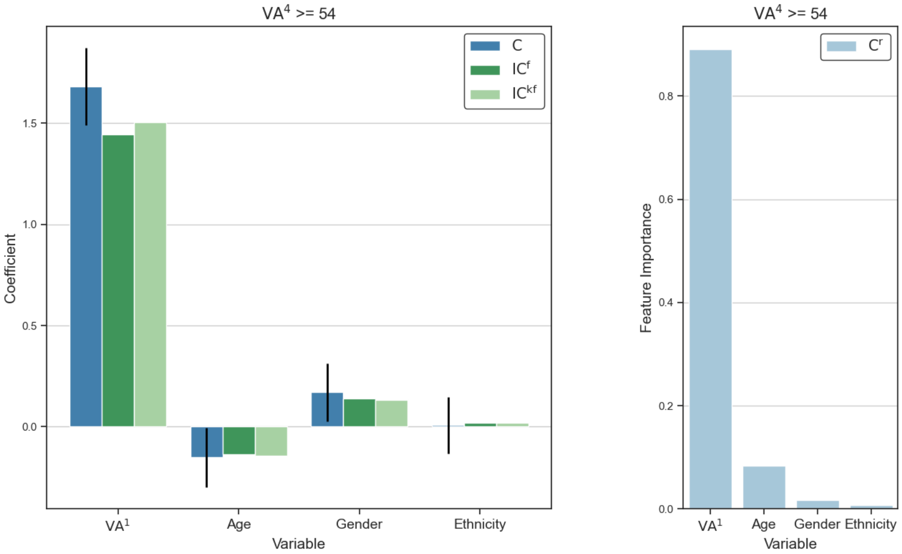


Fig. 3: This figure shows the ROC curves of each model when predicting $y^v$.

than a randomly chosen non-responder, $y^v = -1$. In addition, the sensitivity and specificity 95% confidence intervals of the majority of models indicate that they perform better than a model that gives a random binary prediction to true responders and one that gives a random binary prediction to true non-responders. Note that the specificity 95% confidence interval of the $\mathrm{I^d}$ model includes the value 0.5, and so it does not perform better than a model that gives a random binary prediction to true non-responders. A post-hoc analysis shows that this is due to the optimization of the threshold $T$ that produces the binary predictions: With the default value of 0.5, the sensitivities and specificities are 0.743 and 0.674 with confidence intervals (0.702–0.782) and (0.587–0.753) respectively.

If we now consider table 2 in more detail, we see that the models that use only clinical/demographic features, C and $\mathrm{C^r}$, give AUCs that are 0.096–0.127 higher than those that use only images, I, $\mathrm{I^k}$ and $\mathrm{I^d}$. This implies that the clinical/demographic features are more useful predictors of dichotomized $\mathrm{VA}^4$ than the information in the images that is used by I, $\mathrm{I^k}$ and $\mathrm{I^d}$. We can also see that the best model using only images is the transfer learning model I. Its AUC is 0.023 greater than the transfer learning model using kernel-smoothed image features, $\mathrm{I^k}$, and 0.015 greater than the fully trained model $\mathrm{I^d}$. However, it is interesting to note that a simple ensemble model given by averaging the probabilities output by the $\mathrm{I^d}$ and I models raises the AUC to 0.796. This suggests that the fully trained model has learned complementary information to the transfer learning model for prediction of $y^v$. The highest AUCs are given by the models $\mathrm{IC^f}$ and $\mathrm{IC^{kf}}$, which give AUCs marginally higher than those of models C and $\mathrm{C^r}$.

Figure 4 shows the role that the clinical/demographic variables play in the models which use them as input features. In figure 4a we show the parameter



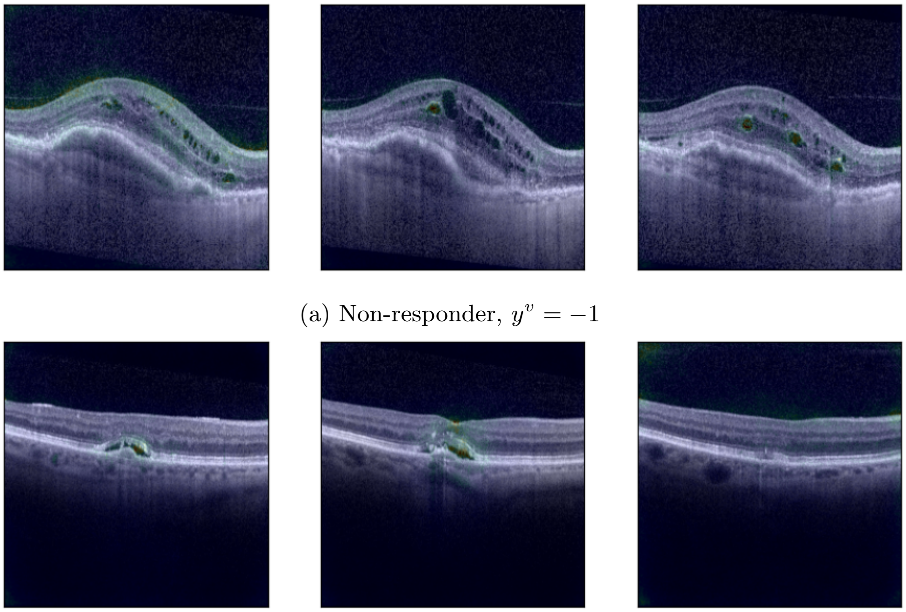
(a)

(b)

Fig. 4: In (a) we show the coefficients of the standardized clinical/demographic feature parameters in the models C, IC$^{\mathrm{f}}$, and IC$^{\mathrm{kf}}$ when predicting $y^v$. 95% confidence intervals are also shown for each parameter for the model C. In (b) we show the clinical/demographic feature importances in the model C$^{\mathrm{r}}$ when predicting $y^v$.

estimates for the standardized clinical/demographic features for models C, IC$^{\mathrm{f}}$, and IC$^{\mathrm{kf}}$. The 95% confidence intervals, based on the standardized normal distribution, are shown for the model C, as reported by the statsmodels package. The parameter estimates indicate positive adjusted associations of VA$^1$, being Female, and being White, and a negative adjusted association of Age, with dichotomized VA$^4$ for all models. For model C, VA$^1$, Age, and being Female have confidence intervals indicating statistically significant adjusted associations with $y^v$. The strong positive adjusted association of VA$^1$ is expected, as a higher VA before treatment would give rise to a higher VA after treatment, even if treatment is ineffective. It should be noted that a variable can be important to a model's predictive accuracy even if it is not statistically significant. Figure 4b shows the feature importances of the random forest model C$^{\mathrm{r}}$ when predicting $y^v$. These importances are the mean decrease in Gini Impurity brought by each feature on the training data, normalized to sum to one over all features. We can see that VA$^1$ and Age are considered to be the most important features for C$^{\mathrm{r}}$.

Figure 5 shows the saliency maps of the model I for prediction of $y^v$. The saliency maps show the magnitude of the gradient of the derivative of the predictive function for a given class, with respect to the pixel values in the preprocessed

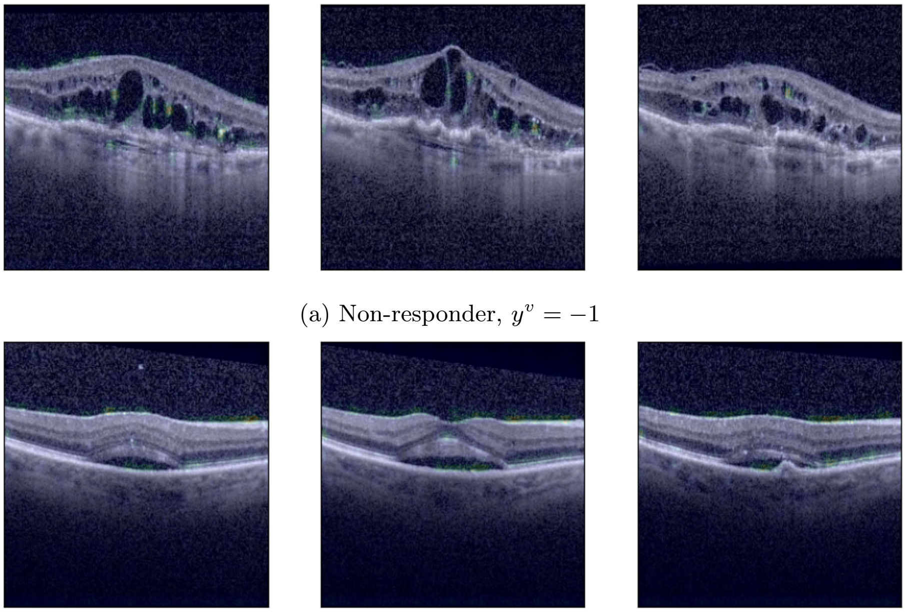


(a) Non-responder, $y^v = -1$

(b) Responder, $y^v = 1$

Fig. 5: This figure shows the saliency maps for two subjects when predicting $y^v$ using model I. The maps are overlayed on the preprocessed OCT scans of the central (foveal) 1mm. These are shown for a non-responder in (a) and a responder in (b).

OCT scans. We use the implementation of 'SmoothGrad' from the tf-keras-vis package, which produces saliency maps by averaging the gradient magnitudes of several noisy versions of the preprocessed OCT scans. In figure 5a we show the saliency map for prediction of a correctly identified non-responder, while figure 5b shows the corresponding map for a correctly identified responder. Note that in each case, the saliency map of a particular b-scan is normalized within its range to ease visual interpretation. If we consider figure 5a, we can see that the most important voxels in a given b-scan (coloured red) are localized to small areas which are sometimes within the regions of intra-retinal fluid. In addition, the voxels of medium importance (coloured green) also overlap with intra-retinal fluid present in the scans. The saliency map in figure 5b seems to indicate that areas in the centre of the retina are important to the prediction of this subject. For comparison, figure 6 shows the saliency maps for a correctly predicted non-responder and responder using the fully trained model $\mathrm{I}^{\mathrm{d}}$. If we consider figure 6a, we can see that voxels of medium importance in a given b-scan are found proximal to regions of intra-retinal fluid. They also overlap with some small areas of hyperintensity. The saliency map in figure 6b indicates areas in the outer retina and at the internal limiting membrane are of medium importance.



(a) Non-responder, $y^v = -1$

(b) Responder, $y^v = 1$

Fig. 6: This figure shows the saliency maps for two subjects when predicting $y^v$ using model $\mathrm{I}^\mathrm{d}$. The maps are overlayed on the preprocessed OCT scans of the central (foveal) 1mm. These are shown for a non-responder in (a) and a responder in (b).

### 4.2 Prediction of $y^{\delta v}$

Table 3: Model performance for prediction of $y^{\delta v}$

| **Model** | **AUC** | **(95% CI)** | **Sensitivity** | **(95% CI)** | **Specificity** | **(95% CI)** |
|---|---|---|---|---|---|---|
| C | 0.743 | (0.703–0.782) | 0.623 | (0.545–0.696) | 0.726 | (0.682–0.766) |
| $\mathrm{C}^\mathrm{r}$ | 0.751 | (0.710–0.790) | 0.760 | (0.688–0.823) | 0.602 | (0.555–0.647) |
| I | 0.571 | (0.518–0.622) | 0.443 | (0.366–0.522) | 0.670 | (0.625–0.714) |
| $\mathrm{I}^\mathrm{k}$ | 0.578 | (0.526–0.630) | 0.479 | (0.401–0.558) | 0.664 | (0.618–0.707) |
| $\mathrm{IC}^\mathrm{f}$ | 0.746 | (0.706–0.784) | 0.617 | (0.538–0.691) | 0.726 | (0.682–0.766) |
| $\mathrm{IC}^\mathrm{kf}$ | 0.749 | (0.709–0.787) | 0.635 | (0.557–0.708) | 0.728 | (0.684–0.768) |
| $\mathrm{I}^\mathrm{d}$ | 0.540 | (0.487–0.593) | 0.491 | (0.413–0.569) | 0.588 | (0.542–0.634) |

Table 3 gives the performance of each model for prediction of $y^{\delta v}$, and figure 7 shows the corresponding ROC curves. All models apart from $\mathrm{I}^\mathrm{d}$ have AUCs with

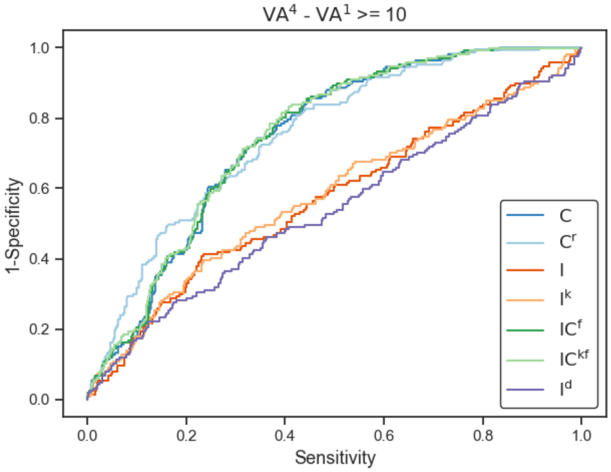
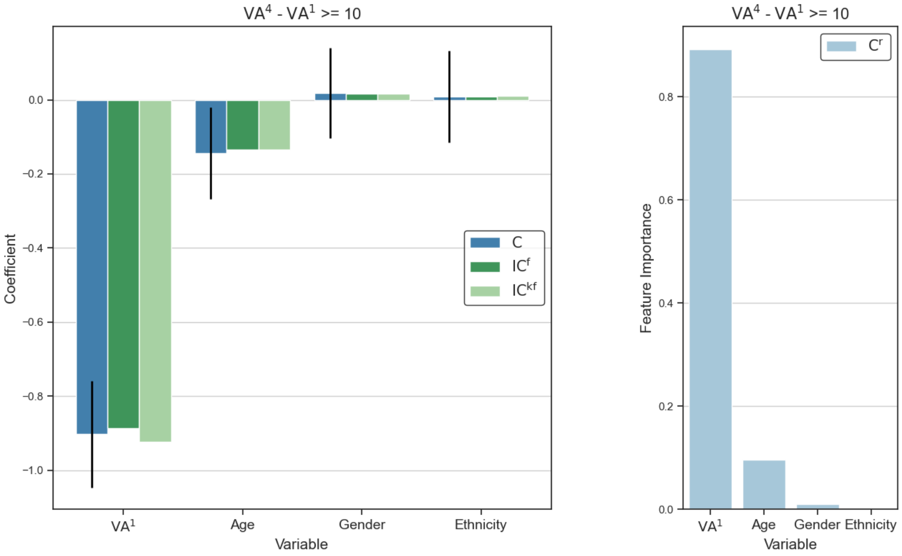


Fig. 7: This figure shows the ROC curves of each model when predicting $y^{\delta v}$.

a 95% confidence interval whose lower bound is greater than 0.5. Hence we have not demonstrated that $\mathrm{I}^{\mathrm{d}}$ performs better than chance with respect to whether a randomly chosen responder, $y^{\delta v} = 1$, is given a higher probability of responding to treatment than a randomly chosen non-responder, $y^{\delta v} = -1$. Therefore we do not consider it to be a useful model for predicting $y^{\delta v}$ and do not discuss it further at this stage. It should also be noted that the confidence intervals for the sensitivity of models I and $\mathrm{I}^{\mathrm{k}}$ include 0.5. This indicates that these models do not perform better than a model that gives a random binary prediction to true responders.

If we now consider table 3 in more detail, we see that the models that use clinical/demographic features have AUCs in the range 0.743–0.751. Moreover, the sensitivities and specificities of each model are very similar, apart from $\mathrm{C}^{\mathrm{r}}$, which has a higher sensitivity and lower specificity than the other models. In particular we can say that there is no obvious improvement in model performance when image-derived features are included with clinical/demographic features over models that use only clinical/demographic features.

Figure 8 shows the role that the clinical/demographic variables play in the models which use them as input features. In figure 8a we show the parameter estimates for the standardized clinical/demographic features for models C, $\mathrm{IC}^{\mathrm{f}}$, and $\mathrm{IC}^{\mathrm{kf}}$, and the 95% confidence intervals for model C. The parameter estimates indicate negative adjusted associations of $\mathrm{VA}^1$, and Age, and positive adjusted associations of being Female, and being White, with dichotomized change in VA for all models. For model C, only $\mathrm{VA}^1$ and Age have confidence intervals indicating statistically significant associations with $y^{\delta v}$. The strong negative adjusted association with $\mathrm{VA}^1$ is the so-called ‘ceiling effect’, in which subjects with smaller pre-treatment VA tend to show larger treatment-related improvements in VA [18]. We would also expect age to be negatively associated with treatment-related gains in VA based on previous literature [18]. Figure 8b shows the feature importances of the random forest model $\mathrm{C}^{\mathrm{r}}$ when predicting $y^{\delta v}$. We can see that $\mathrm{VA}^1$ and Age are considered to be the most important features for this model.

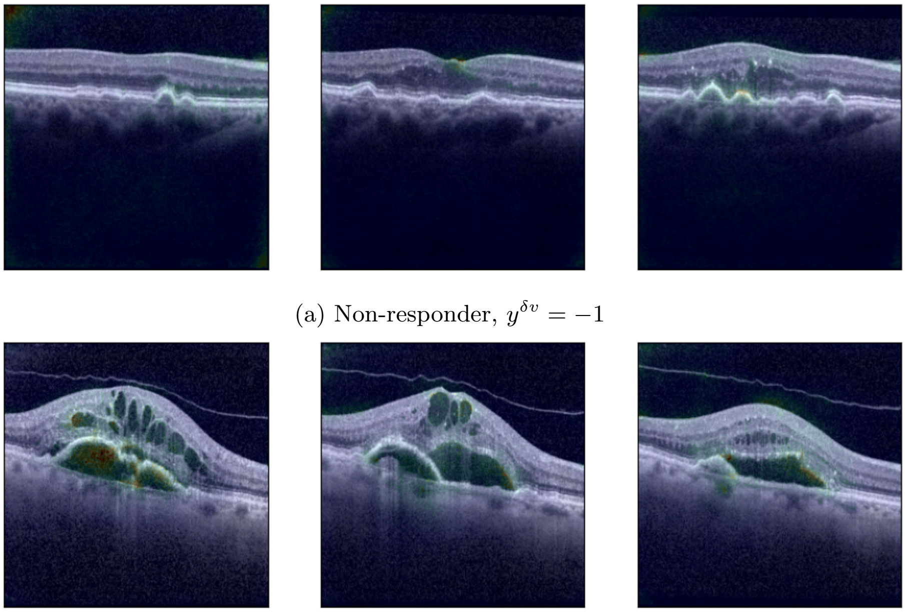


(a)

(b)

Fig. 8: In (a) we show the coefficients of the standardized clinical/demographic feature parameters in the models C, $\mathrm{IC}^{\mathrm{f}}$, and $\mathrm{IC}^{\mathrm{kf}}$ when predicting $y^{\delta v}$. 95% confidence intervals are also shown for each parameter for the model C. In (b) we show the clinical/demographic feature importances in the model $\mathrm{C}^{\mathrm{r}}$ when predicting $y^{\delta v}$.

Figure 9 shows the saliency maps of the model $\mathrm{I}^{\mathrm{k}}$ for prediction of $y^{\delta v}$. In figure 9a we show the saliency map for prediction of a correctly identified non-responder, while figure 9b shows the corresponding map for a correctly identified responder. As before, the saliency map of a particular b-scan is normalized within its range to ease visual interpretation. If we consider figure 9a, we can see that the the voxels of medium importance (coloured green) overlap with some of the pigment epithelial detachment present in the scans. The saliency map in figure 9b shows high (red) and medium importance voxels overlapping with intra-retinal and sub-retinal fluid.



(a) Non-responder, $y^{\delta v} = -1$

(b) Responder, $y^{\delta v} = 1$

Fig. 9: This figure shows the saliency maps for two subjects when predicting $y^{\delta v}$ using model $\mathrm{I}^{\mathrm{k}}$. The maps are overlayed on the preprocessed OCT scans of the central (foveal) 1mm. These are shown for a non-responder in (a) and a responder in (b).

## 5 Discussion

In this work we have used Deep Learning based approaches to predict the response of subjects with treatment-naive nAMD to three monthly doses of intravitreal aflibercept therapy, using OCT and clinical/demographic variables acquired shortly before treatment commenced. When predicting the dichotomized VA 4-10 weeks after the final dose, $y^{v}$, all the transfer learning models and the fully trained Deep Learning model were able to perform better than chance on an unseen test set with respect to the AUC. The transfer learning model that used only image features gave an AUC of 0.787, whereas the logistic regression model using only clinical/demographic features gave an AUC of 0.891. The highest AUC was given by the transfer learning model which used image features and clinical/demographic features, which gave an AUC of 0.901. When predicting the dichotomized change in VA, $y^{\delta v}$, only the models that used just clinical/demographic variables and the transfer learning models were able to perform better than chance. The transfer learning model that used kernel-smoothed image features gave an AUC of 0.578, while the transfer learning model that used kernel-smoothed image features and clinical/demographic features gave an AUC of 0.749. However, the best performing model for this outcome was the random forest model using only clinical/demographic features, which gave an AUC of 0.751.

Given that the models that use only images are the worst performing models for both outcomes, it is interesting to consider if the images are contributing complementary information to the clinical/demographic features in the model predictions. In order to test this, we use a procedure described in [8], which works as follows. Firstly, we fit the outcome of interest using just the clinical/demographic features $c$ as predictors over the test data, and calculate the log-likelihood of the model fit. We then repeat this fit using $c$ and the output of, e.g., the model I on the test data as predictors. These steps can be easily performed using the statsmodels package. A likelihood ratio test of the two model fits then establishes whether the predictions of model I contain different information to the clinical/demographic features $c$ about the outcome. If we do this for outcome $y^v$, we find evidence that the predictions of the models I, $\mathrm{I}^\mathrm{k}$, $\mathrm{I}^\mathrm{d}$ are providing different information about $y^v$ than the clinical/demographic features alone. Thus we have shown it is possible to train Deep Learning Transfer Learning and Fully Trained Deep Learning models using just images that not only perform better than chance for prediction of $y^v$, but also use different information from the clinical variables for its prediction. However, performing the analogous tests for outcome $y^{\delta v}$, does not provide evidence that the predictions of any of the models I, $\mathrm{I}^\mathrm{k}$, $\mathrm{I}^\mathrm{d}$ contain different information to the clinical/demographic features alone. Although I and $\mathrm{I}^\mathrm{k}$ both performed better than chance for prediction of $y^{\delta v}$, this suggests that different modelling approaches should be explored in order to develop predictive models that do not use essentially the same information as that provided by the clinical/demographic variables. Indeed, the best performing model for prediction of $y^{\delta v}$, $\mathrm{C}^\mathrm{r}$ , used only clinical/demographic variables, which further highlights the need for different modelling approaches.

The analysis presented is limited in a number of ways. From a model development perspective, we did not try combining images and clinical/demographic variables in a fully trained model. In [23], they trained such a network to predict an outcome corresponding to $y^{\delta v}$ using OCT images and clinical/demographic variables consisting of best-corrected visual acuity at visit one, age and gender. Although they achieved high accuracy, their study design was markedly different to PRECISE. They aimed to predict a change in VA of more than two lines over a period of a year rather than three months, and some patients in their study were not treatment-naive, unlike the PRECISE study. In future work, we could explore training such models using data that has been acquired with a study design similar to that of PRECISE. We could also consider explicitly including the transfer learning image features in the model to possibly improve model performance. Additional improvements in performance may arise from a careful use of data augmentation of the OCT images when training the Deep Learning models [3, 11]. In experiments we performed using just the training data, we did not find any benefit to incorporating data augmentations based on affine transformations of the OCT images. However, it could be that the use of more flexible elastic transformations, such as those utilized in [24], may improve the predictive performance of the resulting Deep Learning models. Finally, our analysis is limited because the validation was performed with an unseen test set that is a random sample of the complete available data. Although this provides a measure of model performance, it is still considered to be an internal validation. If an independent dataset became available, it could be used to evaluate the models after retraining using the complete PRECISE dataset. This would provide a better assessment of how well these models are likely to perform in a clinical setting.

**Collaborators** The following were collaborators for this multi-site study: Eleni Karatsai, National Institute of Health Research, Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, UK; Geeta Menon, Frimley Health NHS Foundation Trust, Surrey, UK; Benjamin JL Burton, James Paget University Hospitals NHS Foundation Trust, Norfolk, UK; Ian Pearce, The Royal Liverpool and Broadgreen University Hospitals NHS Foundation Trust, Liverpool, UK; Martin Mckibbin, Leeds Teaching Hospitals NHS Trust, Leeds, UK; Ajay Kotagiri, South Tyneside and Sunderland NHS Foundation Trust, Sunderland, UK; James Talks, Newcastle Hospitals NHS Foundation Trust, Newcastle, UK; Anna Grabowska, King's College Hospital NHS Foundation Trust, London, UK; Faruque Ghanchi, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK; Richard Gale, York Teaching Hospital NHS Foundation Trust, York, UK.

**Acknowledgements** The research was supported by the NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology and the NIHR Moorfields Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We thank Asma Burale and Radha Ramakrishnan from National Institute of Health Research, Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, UK. In addition, we would like to thank Adam Dubis from National Institute of Health Research, Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, UK, for his assistance with the data management of this study.

## References

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

2. Balasubramanian, M., Bowd, C., Vizzeri, G., Weinreb, R.N., Zangwill, L.M.: Effect of image quality on tissue thickness measurements obtained with spectral domain-optical coherence tomography. Opt. Express **17**(5), 4019 (Mar 2009). https://doi.org/10.1364/OE.17.004019, https://opg.optica.org/oe/abstract.cfm?uri=oe-17-5-4019
3. Bhatia, K.K., Graham, M.S., Terry, L., Wood, A., Tranos, P., Trikha, S., Jaccard, N.: DISEASE CLASSIFICATION OF MACULAR OPTICAL COHERENCE TOMOGRAPHY SCANS USING DEEP LEARNING SOFTWARE: Validation on Independent, Multicenter Data. Retina **Publish Ahead of Print** (Oct 2019). https://doi.org/10.1097/IAE.0000000000002640, https://journals.lww.com/10.1097/IAE.0000000000002640
4. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth International Group (1984)
5. Chae, B., Jung, J.J., Mrejen, S., Gallego-Pinazo, R., Yannuzzi, N.A., Patel, S.N., Chen, C.Y., Marsiglia, M., Boddu, S., Freund, K.B.: Baseline Predictors for Good Versus Poor Visual Outcomes in the Treatment of Neovascular Age-Related Macular Degeneration With Intravitreal Anti-VEGF Therapy. Invest. Ophthalmol. Vis. Sci. **56**(9), 5040 (Aug 2015). https://doi.org/10.1167/iovs.15-16494, http://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.15-16494
6. Chakravarthy, U., Wong, T.Y., Fletcher, A., Piault, E., Evans, C., Zlateva, G., Buggage, R., Pleil, A., Mitchell, P.: Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis p. 13 (2010)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
8. Dinga, R., Schmaal, L., Penninx, B.W., Veltman, D.J., Marquand, A.F.: Controlling for effects of confounding variables on machine learning predictions. preprint, Bioinformatics (Aug 2020). https://doi.org/10.1101/2020.08.17.255034, http://biorxiv.org/lookup/doi/10.1101/2020.08.17.255034
9. Gill, C.R., Hewitt, C.E., Lightfoot, T., Gale, R.P.: Demographic and Clinical Factors that Influence the Visual Response to Anti-Vascular Endothelial Growth Factor Therapy in Patients with Neovascular Age-Related Macular Degeneration: A Systematic Review. Ophthalmol Ther **9**(4), 725–737 (Dec 2020). https://doi.org/10.1007/s40123-020-00288-0, http://link.springer.com/10.1007/s40123-020-00288-0
10. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. The Journal of Machine Learning Research **12**, 2211–2268 (2011)
11. Gutfleisch, M., Ester, O., Aydin, S., Quassowski, M., Spital, G., Lommatzsch, A., Rothaus, K., Dubis, A.M., Pauleikhoff, D.: Clinically applicable deep learning-based decision aids for treatment of neovascular AMD. Graefes Arch Clin Exp Ophthalmol **260**(7), 2217–2230 (Jul 2022). https://doi.org/10.1007/s00417-022-05565-1, https://link.springer.com/10.1007/s00417-022-05565-1
12. Kiser, A., Mladenovich, D., Eshraghi, F., Bourdeau, D., Dagnelie, G.: Reliability and consistency of visual acuity and contrast sensitivity measures in advanced eye disease. Optometry and Vision Science **82**(11), 946–954 (Nov 2005)
13. Lanzetta, P., Cruess, A.F., Cohen, S.Y., Slakter, J.S., Katz, T., Sowade, O., Zeitz, O., Ahlers, C., Mitchell, P.: Predictors of visual outcomes in patients with neovascular age-related macular degeneration treated with anti-vascular endothelial growth factor therapy: *post hoc* analysis of the VIEW studies. Acta

Ophthalmol **96**(8), e911–e918 (Dec 2018). https://doi.org/10.1111/aos.13751, http://doi.wiley.com/10.1111/aos.13751

14. Lee, C.S., Baughman, D.M., Lee, A.Y.: Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. Ophthalmology Retina **1**(4), 322–327 (Jul 2017). https://doi.org/10.1016/j.oret.2016.12.009, https://linkinghub.elsevier.com/retrieve/pii/S2468653016301749
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
16. Prahs, P., Radeck, V., Mayer, C., Cvetkov, Y., Cvetkova, N., Helbig, H., Märker, D.: OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. Graefes Arch Clin Exp Ophthalmol **256**(1), 91–98 (Jan 2018). https://doi.org/10.1007/s00417-017-3839-y, http://link.springer.com/10.1007/s00417-017-3839-y
17. Russakoff, D.B., Lamin, A., Oakley, J.D., Dubis, A.M., Sivaprasad, S.: Deep Learning for Prediction of AMD Progression: A Pilot Study. Invest. Ophthalmol. Vis. Sci. **60**(2), 712 (Feb 2019). https://doi.org/10.1167/iovs.18-25325, http://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.18-25325
18. Schmidt-Erfurth, U., Waldstein, S.M.: A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Progress in Retinal and Eye Research **50**, 1–24 (Jan 2016). https://doi.org/10.1016/j.preteyeres.2015.07.007, https://linkinghub.elsevier.com/retrieve/pii/S1350946215000671
19. Seabold, S., Perktold, J.: Statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
20. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. Tech. Rep. arXiv:1409.1556, arXiv (Apr 2015), http://arxiv.org/abs/1409.1556, arXiv:1409.1556 [cs]
21. Statistics, L.B., Breiman, L.: Random forests. In: Machine Learning. pp. 5–32 (2001)
22. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
23. Yeh, T.C., Luo, A.C., Deng, Y.S., Lee, Y.H., Chen, S.J., Chang, P.H., Lin, C.J., Tai, M.C., Chou, Y.B.: Prediction of treatment outcome in neovascular age-related macular degeneration using a novel convolutional neural network. Sci Rep **12**(1), 5871 (Dec 2022). https://doi.org/10.1038/s41598-022-09642-7, https://www.nature.com/articles/s41598-022-09642-7
24. Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., Moraes, G., Meyer, C., Wilson, M., Dixon, J., Hughes, C., Rees, G., Khaw, P.T., Karthikesalingam, A., King, D., Hassabis, D., Suleyman, M., Back, T., Ledsam, J.R., Keane, P.A., De Fauw, J.: Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med **26**(6), 892–899 (Jun 2020). https://doi.org/10.1038/s41591-020-0867-7, http://www.nature.com/articles/s41591-020-0867-7