This article has an erratum available at: http://dx.doi.org/10.21037/atm-2021-21 and the article has been update on 2021-06-22 at here.

# Automatic prediction of treatment outcomes in patients with diabetic macular edema using ensemble machine learning

**Baoyi Liu[1]^, Bin Zhang[2], Yijun Hu[3], Dan Cao[1], Dawei Yang[1], Qiaowei Wu[1], Yu Hu[2], Jingwen Yang[2], Qingsheng Peng[1], Manqing Huang[1], Pingting Zhong[1], Xinran Dong[1], Songfu Feng[4], Tao Li[5], Haotian Lin[5], Hongmin Cai[2#], Xiaohong Yang[1#], Honghua Yu[1#]^**

[1]Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China; [2]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China; [3]Aier School of Ophthalmology, Central South University, Changsha, China; [4]Department of Ophthalmology, Zhujiang Hospital of Southern Medical University, Guangzhou, China; [5]Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

*Contributions:* (I) Conception and design: H Cai, X Yang, H Yu; (II) Administrative support: T Li, H Lin, H Cai, X Yang, H Yu; (III) Provision of study materials or patients: S Feng, T Li, H Cai, X Yang, H Yu; (IV) Collection and assembly of data: B Liu, D Cao, D Yang, Q Wu, Q Peng, P Zhong, S Feng; (V) Data analysis and interpretation: B Liu, B Zhang, YHu, H Yu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

*Correspondence to:* Honghua Yu. Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, The Second School of Clinical Medicine, Southern Medical University, No. 106 Zhongshan Er Road, Yuexiu District, Guangzhou 510080, China. Email: yuhonghua@gdph.org.cn; Xiaohong Yang. Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, The Second School of Clinical Medicine, Southern Medical University, No. 106 Zhongshan Er Road, Yuexiu District, Guangzhou 510080, China. Email: syyangxh@scut.edu.cn; Hongmin Cai. School of Computer Science and Engineering, South China University of Technology, Higher Education Mega Center, Panyu District, Guangzhou 510006, China. Email: hmcai@scut.edu.cn.

**Background:** This study aimed to predict the treatment outcomes in patients with diabetic macular edema (DME) after 3 monthly anti-vascular endothelial growth factor (VEGF) injections using machine learning (ML) based on pretreatment optical coherence tomography (OCT) images and clinical variables.

**Methods:** An ensemble ML system consisting of four deep learning (DL) models and five classical machine learning (CML) models was developed to predict the posttreatment central foveal thickness (CFT) and the best-corrected visual acuity (BCVA). A total of 363 OCT images and 7,587 clinical data records from 363 eyes were included in the training set (304 eyes) and external validation set (59 eyes). The DL models were trained using the OCT images, and the CML models were trained using the OCT images features and clinical variables. The predictive posttreatment CFT and BCVA values were compared with true outcomes obtained from the medical records.

**Results:** For CFT prediction, the mean absolute error (MAE), root mean square error (RMSE), and $R^2$ of the best-performing model in the training set was 66.59, 93.73, and 0.71, respectively, with an area under receiver operating characteristic curve (AUC) of 0.90 for distinguishing the eyes with good anatomical response. The MAE, RMSE, and $R^2$ was 68.08, 97.63, and 0.74, respectively, with an AUC of 0.94 in the external validation set. For BCVA prediction, the MAE, RMSE, and $R^2$ of the best-performing model in the training set was 0.19, 0.29, and 0.60, respectively, with an AUC of 0.80 for distinguishing eyes with a good functional response. The external validation achieved a MAE, RMSE, and $R^2$ of 0.13, 0.20, and 0.68, respectively, with an AUC of 0.81.

^ ORCID: Baoyi Liu, 0000-0002-8843-1977; Honghua Yu, 0000-0002-0782-346X.

**Conclusions:** Our ensemble ML system accurately predicted posttreatment CFT and BCVA after anti-VEGF injections in DME patients, and can be used to prospectively assess the efficacy of anti-VEGF therapy in DME patients.

**Keywords:** Machine learning; diabetic macular edema (DME); predictive model; automatic prediction; treatment outcomes

Submitted Feb 08, 2020. Accepted for publication Sep 24, 2020.
doi: 10.21037/atm-20-1431
**View this article at:** http://dx.doi.org/10.21037/atm-20-1431

## Introduction

Diabetic macular edema (DME) is the major cause of vision loss in patients with diabetic retinopathy (DR) (1). The first-line treatment choice in center-involving DME includes three loading doses of anti-vascular endothelial growth factor (VEGF) injections followed by a pro re nata regimen (2). However, approximately 30–50% of DME patients poorly respond to anti-VEGF therapy, and the resolution of macular edema remains transient and partial (2-5). Accurate prediction of treatment response to anti-VEGF therapy in these unresponsive patients on one hand might help vitreoretinal specialists switch to other potentially useful treatments, such as anti-inflammatory therapies, at early stage (4). On the other hand, many DME patients are anxious due to high cost of anti-VEGF therapy (6,7). So, patients are noncompliant to a standard treatment regimen, resulting in the deterioration of prognosis (8). For these noncompliant patients, by improving patient education, reducing psychological stress of the patients and improving their compliance might assist vitreoretinal specialists to predict accuracy on treatment outcomes associated with anti-VEGF therapy. Taken together, a precise prediction of treatment outcomes such as central foveal thickness (CFT) and best-corrected visual acuity (BCVA) after anti-VEGF therapy is considered crucial in treatment planning and delivery.

Optical coherence tomography (OCT) technology has been widely used in diagnosis and follow-up of DME. To better predict the status of DME after anti-VEGF therapy, many OCT parameters of DME that are associated with posttreatment status of DME, such as CFT, preservation of intact inner segment-outer segment junction and external limiting membrane layer, and choroidal thickness have been proposed (9,10). However, these predictions are based on clinician's subjective experience and lacked objective unified standards.

Machine learning (ML), including deep learning (DL) and ensemble learning, has been contemporarily used for the diagnosis and prognosis predictions of many eye diseases (11-14). Numerous efforts have been made recently to accurately detect DME by ML based on OCT images due to continuous increase in the number of patients with DME (15-18). With the development of ML technologies recently, several studies have demonstrated high accuracy in predicting treatment outcomes based on clinical variables (13). With advanced ML algorithms, it has also become possible to predict the posttreatment outcomes of DME based on OCT images and clinical variables.

In the present study, an ensemble ML system consisting of four DL models and five classical ML (CML) models was developed with the aim to predict the posttreatment CFT and BCVA in DME eyes at 1 month after 3 monthly anti-VEGF injections. The four DL models were used to predict the posttreatment CFT and BCVA based on pretreatment OCT images. As the posttreatment outcomes are also associated with many clinical variables (19,20), both OCT image features extracted by the DL models, and the clinical variables obtained by five CML models to predict the posttreatment CFT and BCVA were included. We present the following article in accordance with the TRIPOD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-1431).

## Methods

To predict the posttreatment CFT and BCVA at 1 month after 3 monthly anti-VEGF injections, four DL models and five CML models were developed (*Figure 1*). Three benchmark DL models including AlexNet, Visual Geometry Group (VGG)16, ResNet18 and an ensemble DL scheme of these three models were trained on 304 pretreatment OCT images. For CML models, an ensemble CML scheme was integrated by using four benchmark models, including

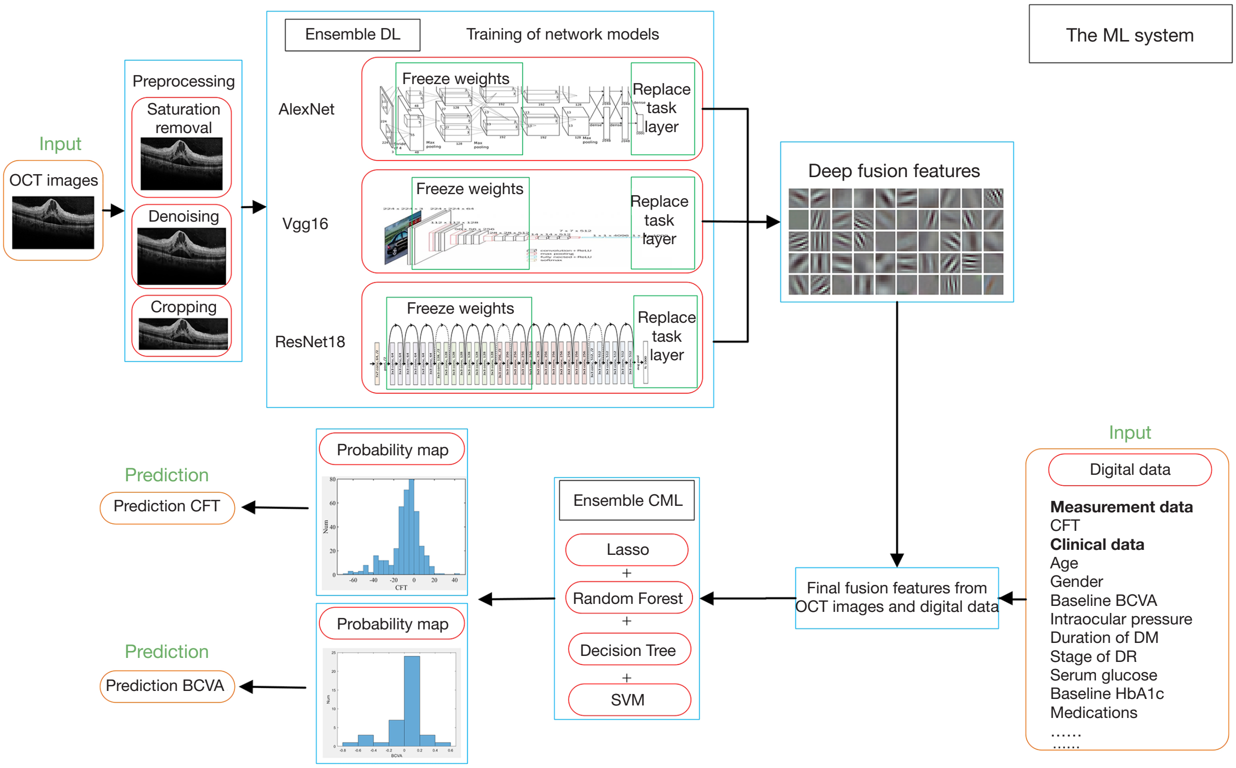

**Figure 1** Demonstration of construction of the ensemble machine learning system. An ensemble DL scheme and three benchmark DL models were used to merge the image features extracted from OCT images. An ensemble CML scheme and four CML models were used to integrate the features obtained from images and digital data to predict CFT and BCVA at 1 month after 3 monthly anti-vascular endothelial growth factor injections in patients with center-involving DME. OCT, optical coherence tomography; DL, deep learning; CNN, convolutional neural networks; VGG, Visual Geometry Group; CFT, central foveal thickness; BCVA, best-corrected visual acuity; DR, diabetic retinopathy; CML, classical machine learning; SVM, support vector machine.

lasso, support vector machine (SVM), decision tree and random forest. These four CML models and the ensemble CML scheme were trained on 15 OCT features extracted by the ensemble DL scheme and 15 clinical variables. The predicted posttreatment CFT and BCVA were compared with the true values obtained from the medical records.

### *Dataset preparation*

From January 2016 to December 2018, 1,243 OCT images and 11,253 longitudinal records from 455 eyes were extracted from two ophthalmic settings, the Department of Ophthalmology, Guangdong Provincial People's Hospital (GDPH) and the Department of Ophthalmology, Zhujiang Hospital of Southern Medical University (ZHSMU), of OCT device and electronic medical records. After data preparation and preprocessing, 304 OCT images and 6,348 clinical data records of 304 eyes obtained from GDPH were used as training set [the mean ± SD age was 57.14±13.90 years, and the baseline CFT and BCVA was 489.13±214.37 µm and 0.79±0.55 logarithm of the minimum angle of resolution (logMAR), respectively]. Meanwhile, 59 OCT images and 1,239 clinical data records of 59 eyes obtained from ZHSMU were used for external validation (the mean ± SD age was 56.81±13.96 years, and the baseline CFT and BCVA was 447.63±186.36 µm and 0.57±0.36 logMAR, respectively). The demographics of all patients included in this study are displayed in Table S1.

All eligible eyes with center-involving DME were included (2). Patients with retinal thickening in the macula involving the central subfield zone of 1 mm in diameter (2) and receiving 3 monthly anti-VEGF injections with complete records of all clinical variables were used and included in the prediction models. Patients with a history of vitrectomy, and any other ocular diseases that might affect ocular circulation (e.g., age-related macular degeneration, glaucoma, retinal artery/vein occlusion, or rhegmatogenous retinal detachment), severe cataracts, or DME previously treated with intravitreal or periocular injections or pan retinal photocoagulation (PRP) within 6 months were excluded. All eligible eyes received 3 monthly anti-VEGF injections after confirmed diagnosis of DME. The anti-VEGF medications used included Lucentis (0.5 mg/injection), aflibercept (2 mg/injection) and conbercept (0.5 mg/injection). After administration of 3 injections, patients were treated according to the recommendations of the latest International Council of Ophthalmology (2). The CFT and BCVA recorded at 1 month after the third anti-VEGF injection were used as the label in the models. This study was conducted according to the Declaration of Helsinki (as revised in 2013) and was approved by the Research Ethics Committee of GDPH (Number: 2016232A). Individual consent for this retrospective analysis was waived.

The images of spectral domain OCT (SD-OCT, Heidelberg Engineering, Heidelberg, Germany) were extracted from the software Heidelberg Eye Explorer version 6.0. A custom of 20°×20° volume acquisition was used to obtain a set of high-speed scans from each eye, wherein 25 horizontal and central vertical cross-sectional B-scan images were obtained, and each composed of 512 A-scans (21). The horizontal image through the fovea was exported to OCT image dataset for manual measurement of CFT according to the simultaneous evaluation of red-free image on the computer monitor of OCT scanner (22). To establish a standardized image format of the dataset for subsequent training and validation, all scans were saved in TIFF format. The 15 clinical variables consisted of pretreatment CFT values from OCT device and 14 variables from medical records included age, sex, eye laterality, pretreatment BCVA measured by decimal charts (converted to logMAR), intraocular pressure, duration of diabetes mellitus (DM), type of DM, stage of DR, previous PRP, random serum glucose, baseline serum hemoglobin A1c (HbA1c) values, concomitant hypertension, and number and medications of intravitreal injections.

### *Preprocessing*

The raw OCT images were first preprocessed to normalize the input data. All saturated pixels with an intensity value of 255 in the OCT images were discarded. The block-matching and 3D filtering (BM3D) (23) method was used for denoising and smoothing the OCT images. The retinal layers were cropped according to the smooth pixel intensity. Finally, the OCT images were resized to 227×227 for AlexNet and 224×224 for both VGG16 and ResNet18.

The clinical variables from the medical records and OCT device were retained and saved into a Microsoft Excel spreadsheet (Microsoft Corporation, Redmond, WA, USA, version 2017) and then loaded into the MATLAB workspace (MathWorks, USA, version 2018A).

### *OCT image features obtained from DL models*

Three dominant convolutional neural networks [CNNs, including AlexNet (24), VGG16 (25) and ResNet18 (26)] and an ensemble DL scheme pretrained on the ImageNet database containing more than a million natural images and a thousand object categories were included in the pretreatment OCT images. The CNNs were fine-tuned via transferred learning, which included freezing weights of the convolution layers that are already optimized to recognize the structures found in images in general at varied depths in the networks, and replacing the deep layers with novel fully connected layers and task layers to be retrained for our new tasks using back propagation algorithm. The architecture on the CNNs was summarized in Table S2. After fine-tuning the three pretrained CNNs, the deep visual features of OCT were extracted from the fully connected layers.

### *Predictive ensemble CML algorithms*

To predict the CFT and BCVA at 1 month after 3 monthly anti-VEGF injections in patients with DME, an ensemble CML scheme was integrated with four benchmark models: lasso (27), SVM (28), decision tree (29) and random forest (30). For these four CML models and ensemble CML scheme mentioned above, in which the inputs included integrated features by combining the deep visual features extracted from the ensemble DL scheme and the clinical variables, were used to predict the output for CFT and BCVA.

The four benchmark models were further integrated to maximize their characteristics and overcome the instability due to limited sample size. The samples were randomly drawn to construct subsets, with each sample being profiled by the integrated features via combining with deep visual features and clinical variables. Four regression models on each subset were trained independently, and then the predicted results were recorded. A probability map of the prediction values was generated, and the mean value of prediction values in the most concentrated regions of the map was employed as the final prediction. Throughout the experiment, 20 subsets were constructed by random sampling from the training set. The four regression models were independently trained on each subset, resulting in 80 predictions for both CFT and BCVA tasks, and a histogram of 80 predictions was created. The three most frequent values were averaged to measure the regression task of CFT prediction. The averaged value for the regression task of BCVA prediction was computed similarly. The three most frequent values were considered reliable and stable predictions among the repetitive experiments.

### *Evaluation of models*

To assess the quality of the predictions per model, three popular evaluation metrics were applied: the mean absolute error (MAE) ($MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-l_i|$), the root mean square error (RMSE) ($RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-l_i)^2}$), and $R^2$ ($R^2=1-\frac{\sum_{i=1}^{n}(y_i-l_i)^2}{\sum_{i=1}^{n}(l_i-\bar{l})^2}$). In the above mathematical formulas, $l$ is the true outcome, $\bar{l}$ is the mean value of $l$. $y$ is the prediction, and $i$ is the number of patients. The lower the MAE and RMSE values, the closer the predicted CFT and BCVA values were to the true value. When the accuracy of each algorithm was described, both the values were usually indicated. The RMSE applied a heavier penalty for the outliers, thereby allowing us to select a more robust algorithm that was specifically helpful if the MAE was comparable between the tested models. Since the CFT and BCVA values were not of the same order of magnitude, their MAEs or RMSEs were considered incomparable. Thus, the normalized $R^2$, which is the coefficient of determination, was used to show the goodness-of-fit of the model (31). As shown in the formula, $R^2$ values closer to 1 indicate a better degree of model fitness.

To test the accuracy of our models for predicting the response of DME eyes towards the 3 monthly anti-VEGF injections, receiver operating characteristic (ROC) curves and area under the curve (AUC) values were calculated as a comprehensive evaluation. Good anatomical and functional responses to anti-VEGF injections were defined as a CFT reduction of more than 50 µm (32), and a BCVA improvement of more than 0.1 logMAR, respectively (33), at 1 month after 3 monthly anti-VEGF injections.

### *Statistical analysis*

#### Internal hold-out validation

To statistically evaluate the results (MAE, RMSE, $R^2$, and AUC), a popular five-fold cross validation (CV) scheme was used on the training dataset. The dataset in this scheme was first randomly split into five independent portions. For each run, four portions were used to train the models, while the last portion was used to evaluate the performance. The experiments were conducted until every portion was tested. The average results after five runs were recorded to measure the overall performance of all models.

#### External validation

In addition to internal validation, OCT images and clinical variables with the same specifications obtained from ZHSMU (not included in internal training and validation), were used to perform an external validation. The results were recorded to evaluate the overall performance of the well-constructed ML system using the best-performing DL model and CML model in the training set.

## Results

The experimental results are summarized in *Table 1*. The detailed degree of fitness between predictions and true outcomes was shown in *Figure 2*, including CFT prediction in the training set (*Figure 2A*) and the external validation (*Figure 2B*), and BCVA prediction in the training set (*Figure 2C*) and the external validation (*Figure 2D*). The AUC of predicting DME patients who responded to anti-VEGF agents was shown in *Figure 3*, including CFT prediction in the training set (*Figure 3A*) and the external validation set (*Figure 3B*), and BCVA prediction in the training set (*Figure 3C*) and the external validation set (*Figure 3D*). To visualize the ensemble ML system's decisions, the weight of different features from clinical variables was shown in *Figure 4*, including CFT prediction task (*Figure 4A*) and BCVA prediction task (*Figure 4B*).

### *Model performance in CFT prediction*

To evaluate the accuracy of DL and CML models, the predicted CFT was compared to the true outcomes obtained from the medical records. In the training set, the performance of DL models was shown to be better with a deeper CNN. The ensemble DL scheme demonstrated the

**Table 1** The accuracy of CFT and BCVA predictions

| Models | CFT, mean (SD) | | | BCVA, mean (SD) | | |
|---|---|---|---|---|---|---|
| | MAE (μm) | RMSE (μm) | $R^2$ | MAE (logMAR) | RMSE (logMAR) | $R^2$ |
| DL models | | | | | | |
| AlexNet | 86.48 (12.62) | 122.46 (22.18) | 0.50 (0.15) | 0.26 (0.03) | 0.38 (0.07) | 0.34 (0.15) |
| Vgg16 | 82.77 (9.52) | 118.44 (17.93) | 0.51 (0.20) | 0.25 (0.03)* | 0.36 (0.06)* | 0.37 (0.13) |
| ResNet18 | 84.85 (11.95) | 116.39 (18.49) | 0.54 (0.13) | 0.28 (0.02) | 0.40 (0.04) | 0.27 (0.12) |
| Ensemble DL scheme | 80.53 (10.45)* | 110.19 (15.37)* | 0.57 (0.20)* | 0.25 (0.04) | 0.36 (0.06)* | 0.39 (0.17)* |
| CML models | | | | | | |
| Lasso | 71.67 (7.50) | 99.07 (11.56) | 0.67 (0.08) | 0.20 (0.03) | 0.31 (0.05) | 0.56 (0.07) |
| SVM | 77.61 (13.68) | 106.82 (18.37) | 0.60 (0.16) | 0.21 (0.03) | 0.34 (0.04) | 0.45 (0.05) |
| Decision tree | 80.69 (5.48) | 116.41 (10.70) | 0.52 (0.17) | 0.24 (0.04) | 0.37 (0.08) | 0.37 (0.26) |
| Random forest | 67.60 (7.51) | 94.19 (13.95) | 0.69 (0.09) | 0.20 (0.02) | 0.30 (0.03) | 0.58 (0.07) |
| Ensemble CML scheme | 66.59 (8.34)* | 93.73 (13.44)* | 0.71 (0.08)* | 0.19 (0.03)* | 0.29 (0.04)* | 0.60 (0.08)* |
| External Validation | | | | | | |
| Ensemble ML System | 68.08 | 97.63 | 0.74 | 0.13 | 0.20 | 0.68 |

Accuracy of CFT and BCVA predictions in DME patients after upload 3 anti-VEGF treatment compared with true outcome. Results were stratified depending on the model performance to the prediction task. The columns represent the CFT and BCVA predictions, which were revealed to the deep-learning and machine-learning models. All mean CFT and BCVA predictions are given with standard deviation. Best results are marked by * in each column. CFT, central foveal thickness; BCVA, best-corrected visual acuity (logMAR); MAE, mean absolute error; RMSE, root mean square error; R2, coefficient of determination; DL, deep learning; Vgg, Visual Geometry Group; CML, classical machine learning; SVM, support vector machine; ML, machine learning.

best performance among all the DL models (MAE, RMSE, and $R^2$ was 80.53, 110.19, and 0.57, respectively), and the ensemble CML scheme demonstrated the best performance among all other CML models (MAE, RMSE, and $R^2$ was 66.59, 93.73, and 0.71, respectively). All the CML models showed a better performance than the DL models. The detailed individual degree of fitness was high as most of the data characterized by the predictions and true outcomes were in between the red line and green line in the plot (*Figure 2A*). For predicting DME patients' response to anti-VEGF therapy, the best AUC was 0.90 in the training set (*Figure 3A*).

In external validation (*Table 1*) using the ensemble DL scheme and the ensemble CML scheme, the MAE, RMSE, and $R^2$ of the CFT task was 68.08, 97.63, and 0.74, respectively, and the mean AUC was 0.94 (*Figure 3B*), which was slightly better than the performance in the training set. The individual details between the prediction and true outcome showed similar trend as that of the training set (*Figure 2B*).

After running all the samples, our ML system identified three features that were considered important and showed stable prognostic performances in predicting the CFT: baseline CFT, type of DM and HbA1c (*Figure 4A*).

### *Model performance in BCVA prediction*

In the training set, the best results in DL models were shown to be MAE, RMSE, and $R^2$ of 0.25, 0.36, and 0.39, respectively (*Table 1*). In contrast to the CFT prediction task, no improvement was observed if the CNN was deeper for the BCVA prediction task. For CML models, the ensemble CML scheme achieved the best performance, with the lowest MAE, RMSE, and the highest $R^2$ (0.19, 0.29, and 0.60, respectively). The performance of CML models was better than that of the DL models, which was similar to that of the CFT prediction task. The degree of fitness was high (*Figure 2C*), and similar to that of the CFT task. In the

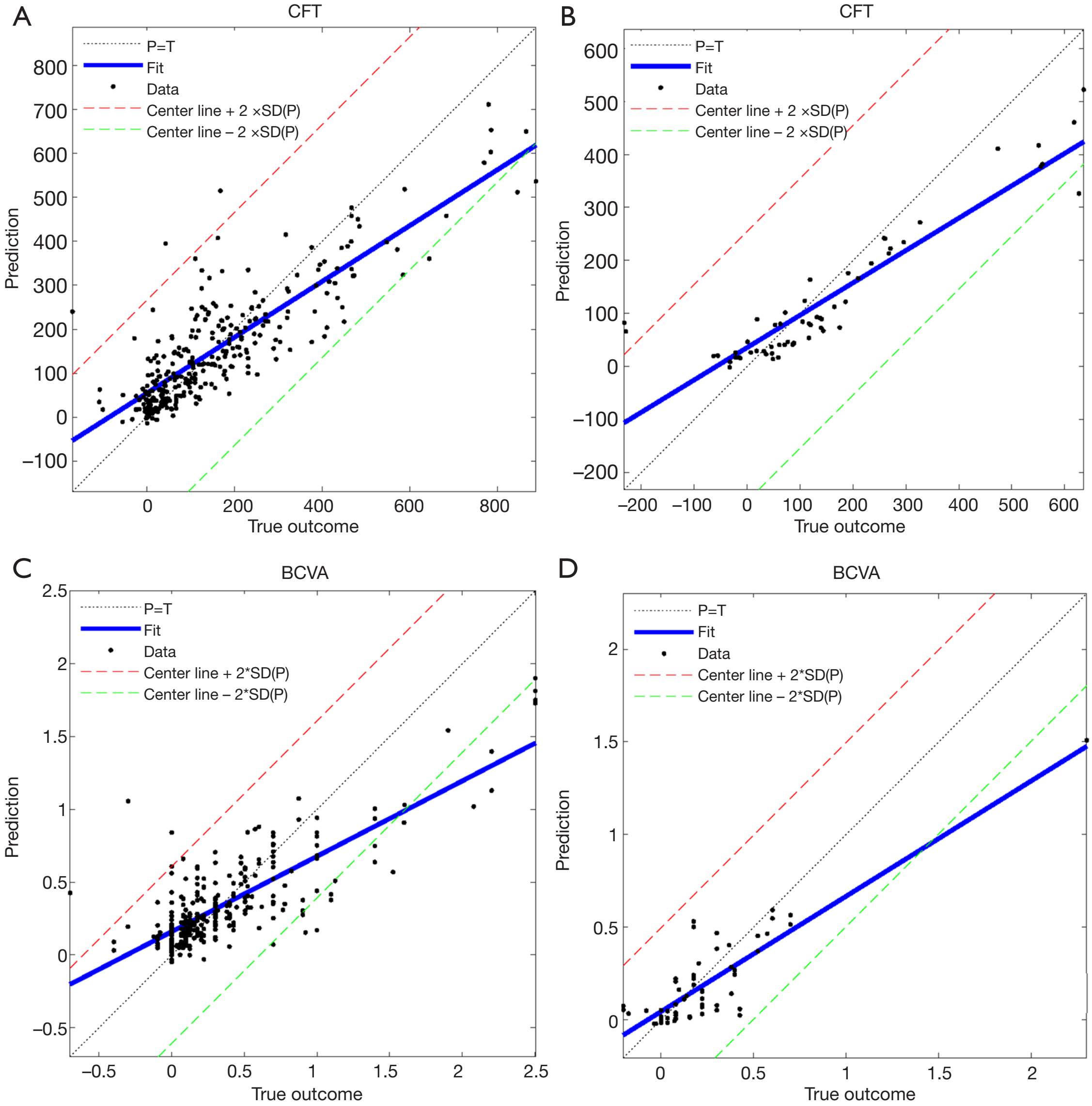

**Figure 2** Degree of fitness between predictions and true outcomes of CFT and BCVA. The blue solid line represents the regression of black dots, which are characterized by the predictions and true outcomes of CFT or BCVA changes. The black dotted line, which represents the centerline, showed that the prediction equals the true outcome. The red or green dotted line is the centerline plus or minus 2 times the standard deviation of the predicted value. The closer the blue solid line to the black dotted line, the higher the fitting degree of the model is. A graph is plotted for each prediction task, including the CFT prediction task in the training set (A) and the external validation (B), and the BCVA prediction task in the training set (C) and the external validation (D). CFT, central foveal thickness; BCVA, best-corrected visual acuity; P, prediction; T, true outcome; SD, standard deviation.

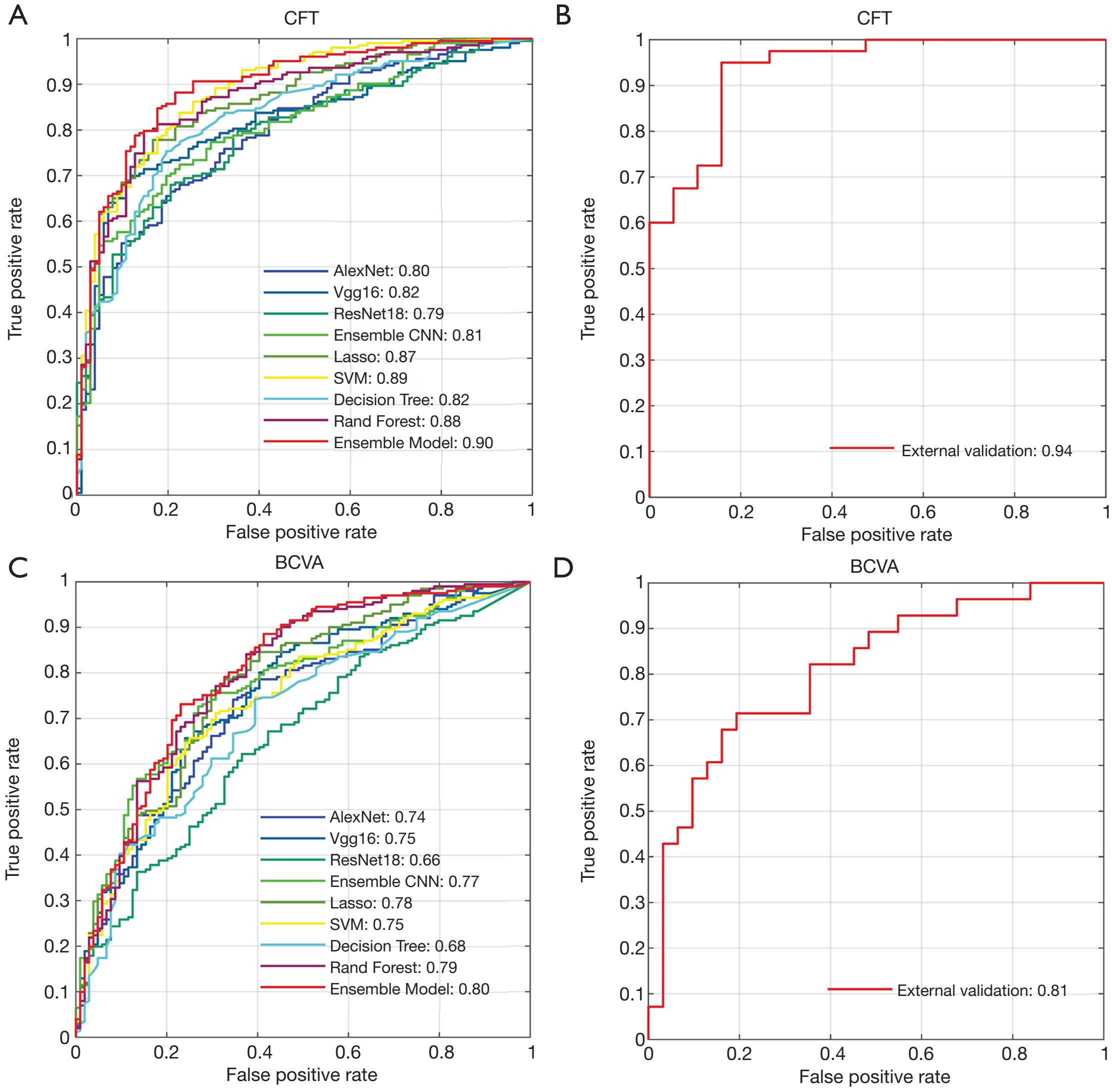

**Figure 3** The AUC of predicting DME patients who respond to anti-VEGF agents. Good anatomical and functional responses to anti-VEGF injections were defined as CFT reduction of more than 50 μm and BCVA improvement of more than 0.1 logMAR, respectively, at 1 month after 3 monthly anti-VEGF injections. The predicted CFT and BCVA values were converted to prediction probabilities using the rule of positive correlation. A series of true positive rates (TPRs) and false positive rates (FPRs) were obtained to form the receiver operating characteristic curve (ROC). The AUC was then calculated as the area of ROC and FPR axis. The TPR is defined as the sensitivity, and the FPR is “1 – specificity”. The AUC of the external validation represents the performance of the ensemble learning system using the ensemble DL scheme and the ensemble CML scheme. A graph is plotted for each prediction task, including the CFT prediction task in the training set (A) and the external validation set (B), and the BCVA prediction task in the training set (C) and the external validation set (D). AUC, area under the receiver operating characteristic curve; CFT, central foveal thickness; BCVA, best-corrected visual acuity.

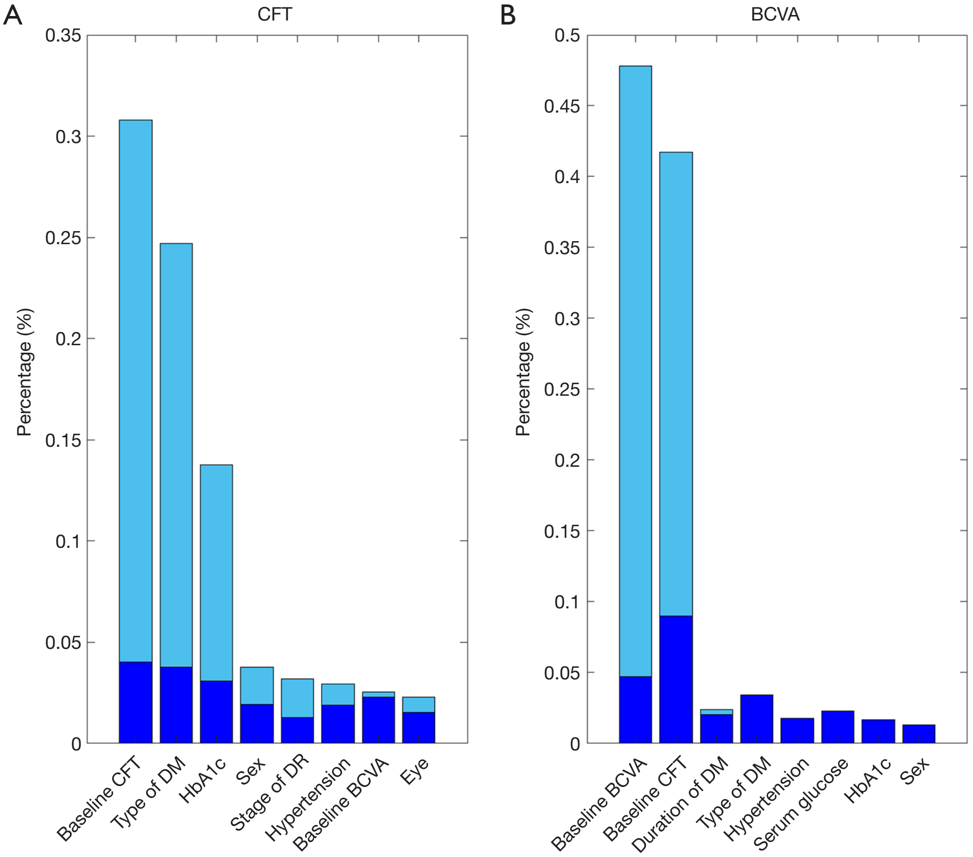

**Figure 4** The weight of different features from clinical variables. Our ensemble learning system identified the weight of different features from twenty clinical variables after running all the samples. The plot demonstrated the weight of different features for CFT prediction task (A) and BCVA prediction task (B). The light blue bar indicates the importance of feature as average for the system on different test runs (5 in total). The higher the light blue bar is, the more important the corresponding feature is for the prediction task. The deep blue bar shows the standard deviation for the model, indicating the stability of the feature. When the light blue bar overlays with the deep blue bar, the feature was shown to be more stable during the test runs. CFT, central foveal thickness; BCVA, best-corrected visual acuity; DM, diabetes mellitus; HbA1c, serum hemoglobin A1c values; DR, diabetic retinopathy; PRP, previous pan retinal photocoagulation.

training set, the mean AUC was 0.80 for predicting DME patients' response to anti-VEGF therapy (*Figure 3C*).

In external validation, the MAE, RMSE, and $R^2$ was 0.13, 0.20, and 0.68, respectively (*Table 1*), and the AUC was 0.81 (*Figure 3D*) using the ensemble DL scheme and the ensemble CML scheme. The individual details between the prediction and true outcome were similar to those of the training set (*Figure 2D*).

Our ML system discovered two features, baseline CFT and baseline BCVA, that exhibited important and showed stable prognostic performance in the BCVA prediction (*Figure 4B*).

## Discussion

The current study developed an ensemble ML system that consisted of four DL models and five CML models to predict the CFT and BCVA values at 1 month after 3 monthly anti-VEGF injections in patients with DME. According to the results of our study, the ensemble ML system accurately predicted the posttreatment CFT and BCVA based on pretreatment OCT images and clinical variables.

Previous studies have shown that prognostic information of patients with DME could be obtained from pretreatment

OCT images. For example, the preservation of intact inner and outer segment junction and external limiting membrane layer before anti-VEGF treatment showed better improvement with CFT and BCVA values after treatment (34-36). In addition, eyes with thicker baseline subfoveal choroidal thickness demonstrated better short-term anatomical and functional responses (10). However, in the past, these pretreatment parameters were used to estimate the likelihood of good or poor treatment response only, and an accurate prediction of posttreatment response was considered not feasible at that time.

So, there is an increasing need to develop precise prognostic predictions in DME patients after anti-VEGF therapy due to increasing number of both DME patients and anti-VEGF medications used (37). With the development of ML, more opportunities than ever were obtained to accurately predict the posttreatment outcomes of anti-VEGF therapy in DME patients. The ML models by using various algorithms can generate more specific and accurate predictions after undergoing training with different morphological, functional, and demographic data.

The posttreatment CFT and BCVA was attempted to predict inpatients with DME based on pretreatment OCT images using the ensemble DL scheme integrated with three benchmark DL models. However, the accuracy of the prediction was shown to be suboptimal. The results were not surprising because the posttreatment outcomes of DME were associated with many factors other than the structure of the retina (38). Therefore, both pretreatment OCT images and fifteen clinical variables that had potential association with treatment outcomes of DME were included into the well-constructed ensemble CML scheme. The model automatically processed vital information (also called "features") for predicting the prognosis and generated the output of CFT and BCVA predictions based on the integration of the above features. Recent studies have shown that the ensemble scheme outperformed each individual alternative, improving the performance obtained by each characterization model separately (39,40). According to a study, an ensemble scheme of four different CNNs was introduced to automatically segment and characterize photoreceptor alteration in macular disease. The results showed that the ensemble scheme outperformed each of its constitutive models with higher Dice coefficient, precision and sensitivity (39). Similarly, the accuracy of the prediction was performed by integrating the ensemble DL scheme and the ensemble CML scheme, which remained satisfactory according to the results of our study. These results indicated that the integration of different sources of information assists in better characterizing the outcome measures, subsequently obtaining more accurate prediction of CFT and BCVA.

The results of our study suggested that for predicting the posttreatment CFT and BCVA values in DME patients, DL models based on OCT images alone were considered insufficient. To achieve a higher accuracy for prediction, other clinical information associated with treatment outcomes should be included in the CML models. This approach is consistent with a well-known fact that the treatment outcomes of DME are affected by many factors, such as clinical variables that reflect the previous physical conditions. The clinical variables selected in our study included duration of diabetes, stage of DR, blood glucose and baseline HbA1c, wherein all these were early phases of DM and DR that showed association with DME treatment outcomes (20,41). Previous studies have indicated that DME patients with a serum HbA1c level of ≤7.0% demonstrated a significant improvement in CFT and BCVA values after undergoing anti-VEGF therapy (19,41). Consistently, the pretreatment CFT, type of DM, and HbA1c showed excellent prognostic performance in CFT prediction according to the weights of different features in our models (*Figure 4*). The second reason as to why the CML models out-performed the DL models in our study was due to the differences in engineering between the two learning models. Usually, DL requires large-scale data to train task-specific feature representation, while CML models require relatively fewer data due to their intrinsic learning strategies. In addition, the DL models directly take raw OCT images as input, while the CML models take image features and clinical data features that have been optimally selected. Therefore, for future predictions, for tasks involving multifactorial diseases such as DME, CML algorithms should be used as they consider multiple factors and are considered effective even when the data are relatively small-scale (empirically <500).

With increasing number of DME patients and anti-VEGF injections worldwide, our ML system was shown to be potentially useful for physicians who oversee the treatment of DME patients. Accurate prediction of two major posttreatment outcomes of DME (CFT and BCVA) provides physicians with valuable information regarding the response of patients to the treatment. This information can help the physician to make better treatment plans for the patients. For patients who are predicted of not benefitting from anti-VEGF treatment, other treatment modalities are recommended, such as anti-inflammatory therapy. On the other hand, personalized treatment response prediction of our system assists physicians to deliver a more customized and precise patient education on treatment outcomes in DME patients. For those patients who are predicted to respond well to anti-VEGF treatment, the promising outcomes regarding anti-VEGF treatment should be emphasized and patients should be encouraged to adhere to the standard treatment regimen. The psychological burden of patients can be reduced in this way, so that patients can better comply with the treatment plan to obtain better treatment effects. More importantly, the predictions from our system were based on common clinical information of DME patients, such as OCT images, with widely used algorithms. This allows more physicians to use the system without the need for additional investment on new ocular examination machines and artificial intelligence computation.

Due to its potential in accurate prediction by combining multiple parameters from a relatively small dataset, our ensemble ML system can also be used in automatic diagnosis or prognosis prediction of many other multifactorial ocular or systemic diseases. Nevertheless, our study has some limitations that should be acknowledged. The outcome prediction in our study was performed in DME eyes treated with anti-VEGF agents, but DME eyes treated with intravitreal injection of other medications such as steroids were not included. However, the ML system that has been developed is expected to predict the outcomes of treatments other than anti-VEGF therapy, if trained properly. A large-sample size and long-term prospective cohort study to optimize the model parameters, and to build a more accurate and stable system for predicting posttreatment outcomes of DME are warranted.

In conclusion, the advanced ML, including DL and ensemble learning, successfully assisted in predicting the postoperative CFT and BCVA values based on pretreatment OCT images and clinical variables in DME patients treated with anti-VEGF. For prediction at 1 month after 3 monthly anti-VEGF injections, the prediction of posttreatment indexes based on our well-constructed ML system might help to better manage the therapy and follow-up in DME patients. Besides, this ML system can be used to predict the treatment outcomes in DME patients, even in a relatively small dataset. Further refinements of the prediction system can be achieved by including larger sample size and conducting longer-term prospective cohort studies in the future.

## Acknowledgments

The authors thank the School of Computer Science and Engineering, South China University of Technology for their technical assistance with the ensemble ML system; and the Department of Ophthalmology, Zhujiang Hospital of Southern Medical University for contributing OCT images and digital clinical data for external validation.

*Funding:* This study was supported by the National Natural Science Foundation of China (81870663 from HY and 61771007 from HC), the National Key R&D Program of China (2018YFC0116500 from HL), the Key R&D Program of Guangdong Province (2018B010109008 from HL), the Outstanding Young Talent Trainee Program of Guangdong Provincial People's Hospital (KJ012019087 from HY), the GDPH Scientific Research Funds for Leading Medical Talents and Distinguished Young Scholars in Guangdong Province (KJ012019457 from HY), the talent introduction fund of Guangdong Provincial People's Hospital (Y012018145 from HY),the Science and Technology Program of Guangzhou (202002030074 from HY), the Science and Technology Planning Project of Guangdong Province (2016A010101013, 2017B020226004 from HC), the Health and Medical Collaborative Innovation Project of Guangzhou City (201803010021 from HC), and the Fundamental Research Fund for the Central Universities (2017ZD051 from HC).

## Footnote

*Reporting Checklist:* The authors have completed the TRIPOD reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-1431

*Data Sharing Statement:* Available at http://dx.doi.org/10.21037/atm-20-1431

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-1431). The authors have no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This retrospective, multicenter study was conducted according to the Declaration of Helsinki (as revised in 2013) and approved by the Research Ethics Committee of GDPH (Number: 2016232A). Individual consent for this retrospective analysis was waived. The patient's personal data have been secured.

## References

1. Ciulla TA, Amador AG, Zinman B. Diabetic retinopathy and diabetic macular edema: pathophysiology, screening, and novel therapies. Diabetes Care 2003;26:2653-64.
2. Wong TY, Sun J, Kawasaki R, et al. Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. Ophthalmology 2018;125:1608-22.
3. Jampol LM, Bressler NM, Glassman AR. Revolution to a new standard treatment of diabetic macular edema. JAMA 2014;311:2269-70.
4. Das A, McGuire PG, Rangasamy S. Diabetic Macular Edema: Pathophysiology and Novel Therapeutic Targets. Ophthalmology 2015;122:1375-94.
5. Gonzalez VH, Campbell J, Holekamp NM, et al. Early and Long-Term Responses to Anti–Vascular Endothelial Growth Factor Therapy in Diabetic Macular Edema: Analysis of Protocol I Data. Am J Ophthalmol 2016;172:72-9.
6. Ho AC, Scott IU, Kim SJ, et al. Anti-vascular endothelial growth factor pharmacotherapy for diabetic macular edema: a report by the American Academy of Ophthalmology. Ophthalmology 2012;119:2179-88.
7. Stewart MW. Anti-VEGF therapy for diabetic macular edema. Curr Diab Rep 2014;14:510.
8. Weiss M, Sim DA, Herold T, et al. Compliance And Adherence Of Patients With Diabetic Macular Edema To Intravitreal Anti–vascular Endothelial Growth Factor Therapy In Daily Practice. Retina 2018;38:2293-300.
9. Brown DM, Nguyen QD, Marcus DM, et al. Long-term outcomes of ranibizumab therapy for diabetic macular edema: the 36-month results from two phase III trials: RISE and RIDE. Ophthalmology 2013;120:2013-22.
10. Rayess N, Rahimy E, Ying GS, et al. Baseline choroidal thickness as a predictor for response to anti-vascular endothelial growth factor therapy in diabetic macular edema. Am J Ophthalmol 2015;159:85-91.e1-3.
11. Togo R, Hirata K, Manabe O, et al. Cardiac sarcoidosis classification with deep convolutional neural network-based features using polar maps. Comput Biol Med 2019;104:81-6.
12. Treder M, Lauermann JL, Eter N. Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier. Graefes Arch Clin Exp Ophthalmol 2018;256:2053-60.
13. Rohm M, Tresp V, Muller M, et al. Predicting Visual Acuity by Using Machine Learning in Patients Treated for Neovascular Age-Related Macular Degeneration. Ophthalmology 2018;125:1028-36.
14. Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng 2017;1:0024.
15. Li Z, Keel S, Liu C, et al. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. Diabetes Care 2018;41:2509-16.
16. Ren F, Cao P, Zhao D, et al. Diabetic macular edema grading in retinal images using vector quantization and semi-supervised learning. Technol Health Care 2018;26:389-97.
17. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. Ophthalmology 2018;125:549-58.
18. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172:1122-31 e9.
19. Matsuda S, Tam T, Singh RP, et al. The impact of metabolic parameters on clinical response to VEGF inhibitors for diabetic macular edema. J Diabetes Complications 2014;28:166-70.
20. Sophie R, Lu N, Campochiaro PA. Predictors of functional and anatomic outcomes in patients with diabetic macular edema treated with ranibizumab. Ophthalmology 2015;122:1395-401.
21. Lee J, Moon BG, Cho AR, et al. Optical coherence tomography angiography of DME and its association with anti-VEGF treatment response. Ophthalmology 2016;123:2368-75.
22. Shimura M, Yasuda K, Yasuda M, et al. Visual outcome

after intravitreal bevacizumab depends on the optical coherence tomographic patterns of patients with diffuse diabetic macular edema. Retina 2013;33:740-7.

23. Dabov K, Foi A, Katkovnik V, et al. editors. Color Image Denoising via Sparse 3D Collaborative Filtering with Grouping Constraint in Luminance-Chrominance Space. ICIP (1); 2007.
24. Krizhevsky A, Sutskever I, Hinton GE. editors. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.
26. He K, Zhang X, Ren S, et al. editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
27. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996:267-88.
28. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273-97.
29. Breiman L. Classification and regression trees. Routledge, 2017.
30. Breiman L. Random forests. Machine Learning 2001;45:5-32.
31. Nakagawa S, Schielzeth H, O'Hara RB. A general and simple method for obtainingR2from generalized linear mixed-effects models. Methods Ecol Evol 2013;4:133-42.
32. Massin P, Bandello F, Garweg JG, et al. Safety and efficacy of ranibizumab in diabetic macular edema (RESOLVE Study): a 12-month, randomized, controlled, double-masked, multicenter phase II study. Diabetes Care 2010;33:2399-405.
33. Diabetic Retinopathy Clinical Research Network, Wells JA, Glassman AR, et al. Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema. N Engl J Med 2015;372:1193-203.
34. Shin HJ, Lee SH, Chung H, et al. Association between photoreceptor integrity and visual outcome in diabetic macular edema. Graefes Arch Clin Exp Ophthalmol 2012;250:61-70.
35. Chung H, Park B, Shin HJ, et al. Correlation of fundus autofluorescence with spectral-domain optical coherence tomography and vision in diabetic macular edema. Ophthalmology 2012;119:1056-65.
36. Maheshwary AS, Oster SF, Yuson RM, et al. The association between percent disruption of the photoreceptor inner segment-outer segment junction and visual acuity in diabetic macular edema. Am J Ophthalmol 2010;150:63-7 e1.
37. Chen SC, Chiu HW, Chen CC, et al. A Novel Machine Learning Algorithm to Automatically Predict Visual Outcomes in Intravitreal Ranibizumab-Treated Patients with Diabetic Macular Edema. J Clin Med 2018;7:475.
38. Ashraf M, Souka A, Adelman R. Predicting outcomes to anti-vascular endothelial growth factor (VEGF) therapy in diabetic macular oedema: a review of the literature. Br J Ophthalmol 2016;100:1596-604.
39. Orlando JI, Gerendas BS, Riedl S, et al. Automated Quantification of Photoreceptor alteration in macular disease using Optical Coherence Tomography and Deep Learning. Sci Rep 2020;10:5619.
40. Orlando JI, Prokofyeva E, Del Fresno M, et al. An ensemble deep learning based approach for red lesion detection in fundus images. Comput Methods Programs Biomed 2018;153:115-27.
41. Bressler SB, Odia I, Maguire MG, et al. Factors Associated With Visual Acuity and Central Subfield Thickness Changes When Treating Diabetic Macular Edema With Anti-Vascular Endothelial Growth Factor Therapy: An Exploratory Analysis of the Protocol T Randomized Clinical Trial. JAMA Ophthalmol 2019;137:382-9.

**Cite this article as:** Liu B, Zhang B, Hu Y, Cao D, Yang D, Wu Q, Hu Y, Yang J, Peng Q, Huang M, Zhong P, Dong X, Feng S, Li T, Lin H, Cai H, Yang X, Yu H. Automatic prediction of treatment outcomes in patients with diabetic macular edema using ensemble machine learning. Ann Transl Med 2021;9(1):43. doi: 10.21037/atm-20-1431

## Visualization method of ensemble DL scheme

To visualize the critical components in OCT images that are highly correlated with CFT and BCVA prediction, a popular occlusion test was used to interpret the results and increase model transparency. A blank 100×100 pixel box was systematically moved across every possible position in the image and the probabilities of the prediction were recorded. The highest drop in the probability represents the part of the OCT image that is most critical for accurate classification (shown as the red part in the Figure S1). Furthermore, whether the identified regions by occlusion test were the most clinically significant areas of predictive basis in DME eyes was further verified by our retinal specialists (YH, DC, and HY).

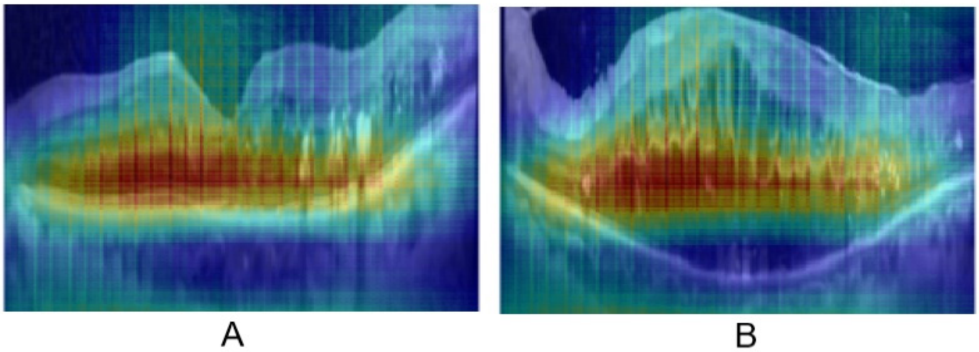

**Figure S1** Occlusion test for visualization of ensemble deep learning scheme. Occlusion test successfully identified the predictive basis in the OCT images from different patterns of DME eyes. An occlusion map was generated by convolving an occluding kernel across the input image. The occlusion map is created after prediction by assigning the SoftMax probability of the correct label to each occluded area. The occlusion map could then be superimposed on the input image to represent the critical components in OCT images that showed highly correlation with the accurate prediction of CFT and BCVA in DME patients. The red part represents high correlation, while the blue part represents low correlation.

**Table S1** Patient demographics

| Variable | Training set | Validation set |
|---|---|---|
| No. of patients [female] | 208 [143] | 41 [22] |
| No. of eyes | 304 | 59 |
| Age, mean (SD), years | 57.14 (13.90) | 56.81 (13.96) |
| Preoperative CFT, mean (SD), μm | 489.13 (214.37) | 447.63 (186.36) |
| Postoperative CFT, mean (SD), μm | 334.15 (137.53) | 303.54 (92.47) |
| No.(percentage) of eyes responding in CFT | 202 (66.45) | 40 (67.80) |
| Preoperative BCVA, mean (SD) | 0.79 (0.55) | 0.57 (0.36) |
| Postoperative BCVA, mean (SD) | 0.44 (0.41) | 0.32 (0.28) |
| No.(percentage) of eyes responding in BCVA | 200 (65.79) | 37 (62.71) |

No., number; SD, standard deviation; CFT, central foveal thickness; BCVA, the best-corrected visual acuity [in the logarithm of minimum angle of resolution (logMAR) unit].

**Table S2** The properties of applied CNNs

| Networks | AlexNet | Vgg16 | ResNet18 |
|---|---|---|---|
| Depth | 8 | 16 | 18 |
| Parameters (millions) | 61.0 | 138 | 11.7 |
| Image input size | 227-227-3 | 224-224-3 | 224-224-3 |

Depth means the largest number of fully-connected layers or sequential convolutional layers on a path from the input layer to output layer. Parameters were defined as the number of weights in the networks. The image input size means the required sizes of input images, in which 3 is the number of color channels and 227 or 224 is the number of pixels. VGG, visual geometry group.