Peer Review File

Article information: http://dx.doi.org/10.21037/atm-20-1789

Reviewer:

Major concerns

Comments 1: No details are given about development of the DL method, such as training parameters (optimization, batch size…)? How many nodes are used in the last layer to generate the predictions?). All these decisions can influence the final performance of the model as well as the quality of the proposed method.

Reply 1: The details about development of the DL method have been added into the Method section, as follows:
The training started with multiple iterations with a batch size of 16 images, with the initial learning rate of 0.001 and stopped at 15 epochs. For each training iteration, a cross-entropy loss function was used as the objective loss function to update the optimization parameters, and a stochastic gradient descent (SGD) algorithm with Nesterov momentum term was used to optimize a pre-defined loss function to train neuron weights via backpropagation. At every epoch, the performance of the convolutional neural network (CNN) was assessed using the validation dataset. The VGG 16 model is a binary classification model output by the Softmax classifier, and 2 nodes are used in the last layer to generate the predictions.

Change in the text 1: We have revised the details of the DL method as suggested (see Page 7, line 154-161).

Comments 2: While the authors include a table to illustrate patient demographics in the internal and external validation sets, the prevalence of close and open macular holes (labels used to train and validate the model) in the available datasets is not mentioned; they can only be intuited from the confusion matrices. Such prevalence can also condition the development and performance of a DL method.

Reply 2: The prevalence of close macular hole in the internal validation set and the external validation set was 68.80% and 66.67%, respectively.

Change in the text 2: We have added the prevalence of close and open macular holes in Table 1.

Comments 3: I have some concerns regarding the data partition and use.

a. The authors mention the internal set is divided into 10 independent folds. This partition should be done at the patient-level in order to have fully independent folds. There is more than one image per eye in many cases, plus eyes could belong to the same patient in some cases. Can the authors confirm this? In that case, it is necessary to indicate it in the paper (how the partition was done).

b. The authors do not mention that any of the 10 folds is used for testing. I understand they are only used for training and validation. However, the validation set is normally used for model tuning and optimization, therefore, that set is not completely independent from model development, and results reported on that set cannot be considered as an independent validation.

Reply 3a: Thank you for the insightful and helpful comments. We have modified the Method section with more details as follows:

In the training dataset, there were 96.88% of patients with more than one OCT image per eye, with an average of 3.6 OCT images per eye. To determine the most appropriate hyperparameters for our final model, the popular ten-fold cross-validation scheme was used on the training dataset that was divided into ten portions randomly. To have fully independent folds, we first used python code to read the labels and corresponding index positions of all images. Then, we used the code to remove the repeated names of different index position images of the same patient (in other words, only one patient name was kept in several images of the same patient). Therefore, the partition was carried out at the patient-level, and the original index position images belonging to the same patient were divided into the same partition to ensure the completely independent folds.

Change in the text 3a: We have modified the text in the Method section (see Page 7-8, line 162-170).

Reply 3b: Cross-validation is indeed not a completely independent validation, which is mainly used for model tuning and optimization. We have modified the Method section as follows:

In each run of the training and internal validation, nine portions of the dataset were employed to train the DL model and the other one was used for model testing to facilitate parameter selection and tuning. The experiments conducted until each portion was tested. Using the above-chosen hyperparameters, we re-trained the DL system using the entire training dataset and evaluated its predictive performance on our independent external validation dataset, whose data were not involved in the development of the DL-based method.

Change in the text 3b: We have modified the text in the Method section (see Page 8, line 170-175).

Comments 4: The authors mention the second goal of the paper, at the end of the Introduction section, that they demonstrate the areas correlated to the predictions in the input images. It is positive that the authors apply one of the available visual attribution methods to obtain such additional information, however, as correctly not included in the Abstract, this aspect cannot be considered a main goal of the paper, since the maps are not validated for detection, and are only visually examined

in 2 images, providing coarse visualizations.

Reply 4: We have removed the relevant content from the Introduction section.

Change in the text 4: We have removed the relevant content from the Introduction section as advised.

Comments 5: The discussion section is messy and does not contain an in-depth analysis of the proposed method and the obtained results. It contains descriptions of related clinical interventions, which would belong to the Introduction, as well as other relevant DL methods which were also not described in the Introduction. The Introduction lacks, therefore, this information. While it is a positive comparison with other state-of-the-art methods in the Discussion, it does not make sense to compare numerically the performance with that of methods developed for other image modalities and/or other tasks, such as prediction of DR progression in color fundus images.

Reply 5: Thank you for the insightful comments.
Firstly, we have added discussion about the in-depth analysis of the proposed method and the obtained results in the Discussion as follows (Page 10-12, line 231-274):
Secondly, we have deleted the descriptions of related clinical interventions and other relevant DL methods in the Discussion section, and added them into the Introduction section (Page 4, line 84-87; Page 5, line 97-102).

Thirdly, we have deleted the numerical comparison with the performance of methods developed for other image modalities and/or other tasks prediction, including prediction of DR progression in color fundus images, in the Discussion section.

Change in the text 5: We have added more content into the Discussion section (see Page 10-11, line 231-274). We also have moved some content to the Introduction section (see Page 4, line 84-87; Page 5, line 97-102).

Comments 6: The use of English should be checked along with the paper.

Reply 6: The English in the manuscript has been carefully polished.

Minor concerns

Comments 1: Certain technical terminologies should be checked so as to be consistent within the

paper and with the terms used by the scientific community (such as “hot maps”).

Reply 1: We appreciate your attention to details. We have already checked the technical terminologies and changed “hot maps” to “heatmaps” as suggested. We also have changed the abbreviation “FTMH” to be “MH” so that it is consistent within the paper.

Change in the text 1: We have changed “hot maps” to “heatmaps” (see Page 9, line 205; Page 9, line 215) and the abbreviation “FTMH” to be “MH” (see Page 3, line 53; Page 4, line 80; Page 5, line 104).

Comments 2: I think only the B-scan containing the macular center is used as input to the model. However, this is not completely clear and should be highlighted in the sections about Data and Model development.

Reply 2: We have changed the text in the Method section to make it clear about the input of the models following:
The VGG16 consisting of 16 convolutional layers was used as the benchmark DL-based model in our whole experiments, in which the inputs were the OCT images (B-scans containing the macular center) and the outputs were the status of MH after VILMP.
Change in the text 2: We have revised the text in the Method section (see Page 7, line 151-152).

Comments 3: The surgical process related to the data and task is now included within the Methods section, while such surgical procedure is not part of the proposed method. This information belongs to the Introduction.

Reply 3: We have deleted the description of the surgical process in the Method section and briefly described it in the Introduction section.

Change in the text 3: We have modified our text as suggested (see Page 4, line 80-82).

Comments 4: The goal and information of the “subgroup analysis” regarding the provided metrics are not clear. The authors already provide values for AUC, sensitivity (SE), and specificity (SP), which indicate global performance in terms of positive and negative samples.

Reply 4: The statement of the “subgroup analysis” has been deleted in the Result section.

Change in the text 4: We have deleted the statement of the “subgroup analysis” in the Result section

as advised.

Comments 5: It is necessary to define how SE and SP are computed from the ROC curve, i.e., which point is considered as an optimal operation point?

Reply 5: We have added the details about the ROC curve analysis in the Method section as follows: A series of true positive rates (TPRs) and false positive rates (FPRs) were obtained to form ROC curves. The TPR is also known as sensitivity, and the FPR results from subtracting the specificity value from 1. The optimal cut-off point was obtained by using the highest Youden's index (sensitivity+specificity−1), and the corresponding optimal sensitivity and specificity values are recorded.

Change in the text 5: We have modified our text as advised (see Page 8, line 182-186).