Peer Review File

Article information: http://dx.doi.org/10.21037/atm-20-6635

Reviewer A

Comment 1: *Although the successful demonstration may contribute to lens opacity detection in clinical practices, the current form had several major concerns and made its scientific contribution unjustified.*

Reply 1: Thanks so much for your agreement on our contribution to lens opacity detection in clinical practices. We appreciate your technical criticisms as well. We conducted more supplementary experiments to further verify the effectiveness of our CCNN-Ensemble for lens opacity detection. We hope that the detailed comparative experiments and interpretations could make the scientific contribution of our study justified (see our later response).

Reviewer B

Comment 1: *This study present of an ensemble technique of conventional CNN models to diagnose pediatric cataract using slit-lamp images. The proposed method achieved improved performance in classification with the use of the cost-sensitive ensemble method. The research group has performed great works for AI & ophthalmology and I have admired their recent researches. I have the following specific comments.*

Reply 1: Thanks so much for your agreement on the methodology and approach. We appreciate your technical criticisms as well, which is constructive for promoting our study. We have made revisions according to your comments and highlighted corresponding changes in the revised manuscript.

Comment 2: *There is no related literature review about ensemble techniques for deep learning in the introduction.*

Reply 2: Thank you for your advice. We carefully reviewed and added the related literatures as follows：Recently, convolutional neural networks (CNNs) (26-28) and ensemble learning methods (29-32) based on CNNs showed great promise in the automatic diagnosis of extensive diseases based on medical images, among which, the voting, averaging, and batch random selection were common ensemble techniques.

Changes in the text: Please see Page 6, Line 124-127.

Comment 3: *The authors should implement other ensemble techniques such as voting or averaging and compare them with the proposed method. Because of the batch random selection, 3 or 10 same architecture (for example, 10 GoogleNet or 3 ResNet50, I think this is the most common ensemble technique) would be better than the ensemble models using heterogenous* *models. Please perform more experiments to show that the proposed method is better than the conventional ensemble methods.*

Reply 3: Many thanks for your constructive suggestions. We agree with the reviewer that the voting and averaging techniques are common ensemble learning methods, and the batch random selection is a common technique for building basic classifiers of ensemble learning. Following your suggestions, we further conducted comparative experiments including other two ensemble learning methods (AVE-Ensemble and AVE-BRS-3ResNet) and three single-classifier CNNs (AlexNet, GoogLeNet, and ResNet50). The Ave-Ensemble represents an ensemble learning with an averaging technique, which calculates the averages of the probabilities for AlexNet, GoogLeNet, and ResNet50 to obtain the final classification result. The Ave-BRS-3ResNet denotes the ensemble learning of three ResNet50 architectures with batch random selection and averaging techniques.

From the experiment results of opacity area grading (Table 3), we obtained three meaningful conclusions. First, the performance of three ensemble learning methods was superior to those of the three single-classifier CNNs. Compared with the best single-classifier ResNet50, the accuracy, specificity, and sensitivity of the CCNN-Ensemble were improved by 3.45%, 2.54%, and 4.62%, respectively. Second, the performance of the Ave-Ensemble and the Ave-BRS-3ResNet is comparable. Third, the performance of CCNN-Ensemble was superior to those of the Ave-Ensemble and the Ave-BRS-3ResNet methods. It was worth to note that the sensitivity of the CCNN-Ensemble was improved by 3.08% when compared to that of the Ave-Ensemble method. Similar conclusions were obtained on the grading of opacity density and location (Supplementary Table S3–S4). Accordingly, we have added these comparative experiments and the detailed discussion in the manuscript.

Table 3. Performance comparison of CCNN-Ensemble with single-classifier CNNs and conventional ensemble learning methods based on CNNs in opacity area grading.

Method	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC
AlexNet	87.48(2.94) §	88.64(3.15)	86.15(3.64)	85.25(3.12)	86.16(3.14)	90.41(2.46)
GoogLeNet	88.16(2.63)	89.30(2.49)	86.67(5.25)	85.75(3.20)	87.94(2.94)	92.84(2.04)
ResNet50	88.68(1.25)	89.46(3.27)	87.69(2.74)	86.40(1.43)	88.54(1.91)	93.60(1.83)
Ave-Ensemble	90.28(1.61)	90.46(2.40)	89.23(2.65)	87.89(1.67)	89.83(1.41)	94.87(1.72)
Ave-BRS-3ResNet	89.50(1.84)	90.12(2.68)	88.72(2.92)	87.38(1.94)	89.39(1.58)	94.02(1.93)
CCNN-Ensemble	***92.13(1.21)***	***92.00(2.07)***	***92.31(2.56)***	***90.68(1.42)***	***92.14(1.25)***	***97.76(0.81)***

Changes in the text: Please see (Page 9, Line 187-143; Page 15, Line 314-316; Page 17, Line 352-369; Page 21, Line 448-460; Table 3; Supplementary Table S3–S4).

Comment 4: *The readers will want to know the performance of the single deep learning architecture without ensemble.*

Reply 4: Many thanks for your constructive suggestions. We performed additional experiments as you advised. Three single-classifier CNNs (AlexNet, GoogLeNet, and ResNet50) were used to compare with the CCNN-Ensemble. From the experiment results (Table 3), the performance of the three ensemble learning methods is superior to those of the three single-classifier CNNs. Compared with the best single-classifier ResNet50, the accuracy, specificity, and sensitivity of the CCNN-Ensemble were improved by 3.45%, 2.54%, and 4.62% in the grading of opacity area, respectively. Similar conclusions were obtained on the grading of opacity density and location (Supplementary Table S3–S4). Accordingly, we have added these comparative experiments and the detailed discussion in the manuscript.

Changes in the text: Please see (Page 9, Line 187-143; Page 15, Line 314-316; Page 17, Line 352-369; Page 21, Line 448-460; Table 3; Supplementary Table S3–S4).

Comment 5: *The detailed conditions of slit-lamp images are needed. The description about slit-beam width, camera specification, and cropping method will clarify the data collecting process.*

Reply 5: Many thanks for your suggestions. In total, there were 470 individuals in the training and validation datasets and 132 individuals in the independent testing dataset. All individuals underwent the examination of slit lamp-adapted anterior segmental photography (BX900; Haag-Streit AG, Koniz, Switzerland). The slit-beam width was settled in a narrow range (1 to 2 mm). Auto-cropping was employed to minimize the noise around the lens based on our previous method (7, 9). To clarify this, we have added these information in the manuscript.

Changes in the text: Please see Page 7-8, Line 153-157; Page 9, Line 178-179 and references (7, 9).

Comment 6: *Mean age of the subjects is needed to clarify the dataset.*

Reply 6: Many thanks for your advice. The age of the subjects in training, validation and independent testing datasets is 18.96 ± 10.61 months (mean±SD). To clarify this, we have added this information in the manuscript.

Changes in the text: Please see Page 8, Line 157-159.

Comment 7: *The input sizes of GoogleNet, AlexNet, and ResNet may be different. Did the authors resize the images or modify the input tensors?*

Reply 7: Thank you for your question. For a fair comparison, after automatically cropping the lens region, all images were resized to 256*256 pixels and input into the three single-classifier CNNs (GoogleNet, AlexNet, and ResNet) and ensemble learning methods. We have added this statement in the manuscript.

Changes in the text: Please see Page 14, Line 303-305.

Comment 8: *There is no comment about explanability. The authors should show the heatmap from the Grad-CAM technique. If the proposed method is used in the clinics, how will the authors select the heatmap from the different deep learning architectures?*

Reply 8: We appreciate your constructive suggestion and question! We agree with the reviewer that the interpretability analysis for CCNN-Ensemble is necessary. We agree that the Gradient-weighted Class Activation Mapping (Grad-CAM) (46) is an effective explainable technique for CNN-based models, which utilized the gradients of any target concept flowing into the last convolutional layer to produce a localization map highlighting remarkable regions in the image for predicting the concept. Following your suggestion, the Grad-CAM visualization technique was employed to generate the heatmaps for highlighting the disease-related regions on which the CCNN-Ensemble model focused most. Using the Grad-CAM technique, three heatmaps can be obtained simultaneously via the CCNN-Ensemble, which are associated with the Alexnet, GoogLeNet, and ResNet50, respectively. By analyzing the heatmap shown in Fig. 6, it finds that all these three classifiers can capture the location of lens opacity, although the highlighted areas are slightly different. Accordingly, we have added the detailed information for the Grad-CAM and the experiment result in the manuscript

Changes in the text: Please see (Page 14, Line 288-295; Page 18-19, Line 393-399; Page 21, Line 455-460; Page34, Line 711-716 and Figure 6).

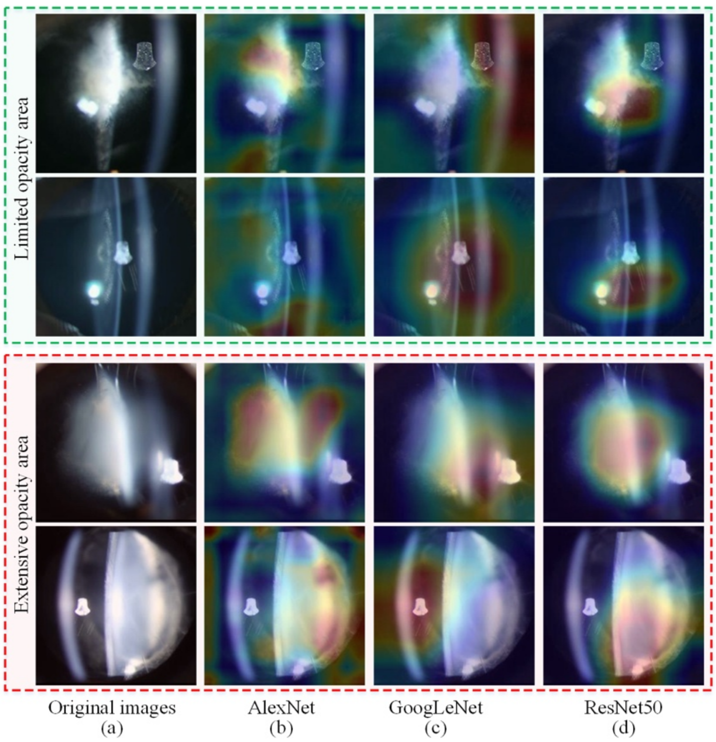

Fig 6. The representative heatmaps of CCNN-Ensemble in opacity area grading using Grad-CAM. (a) The original slit-illumination images. **(b)–(d)** The visualization heatmaps generated from Alexnet, GoogLeNet, and ResNet50 in the CCNN-Ensemble method. The upper two rows indicate negative samples with limited opacity area, and the lower two rows represent positive samples with extensive opacity area. Footnote: Grad-CAM: Gradient-weighted Class Activation Mapping.

Comment 9: *Does the dataset include the normal lens group?*

Reply 9: Thank you so much for your question. In our previous studies, we have developed a deep learning system that could detect cataract with high accuracy from slit-lamp images (7-9). In this study, we provided an effective CCNN-Ensemble method for grading lens opacity of cataract cases. The normal lens group of slit-lamp images was not included in this study. For clarity, we modified this statement in the manuscript.

Changes in the text: Please see Page 5, Line 93-97 and references (7-9).