Original Article

Automatic classification of heterogeneous slit-illumination images using an ensemble of cost-sensitive convolutional neural networks

Jiewei Jiang[1#], Liming Wang[2#], Haoran Fu[2#], Erping Long[3], Yibin Sun[1], Ruiyang Li[3], Zhongwen Li[3], Mingmin Zhu[4], Zhenzhen Liu[3], Jingjing Chen[3], Zhuoling Lin[3], Xiaohang Wu[3], Dongni Wang[3], Xiyang Liu[2], Haotian Lin[3]

[1]School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China; [2]School of Computer Science and Technology, Xidian University, Xi'an, China; [3]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China; [4]School of Mathematics and Statistics, Xidian University, Xi'an, China

Contributions: (I) Conception and design: J Jiang, L Wang, H Fu, H Lin, X Liu; (II) Administrative support: H Lin; (III) Provision of study materials or patients: E Long, R Li, Z Li, J Chen; (IV) Collection and assembly of data: J Jiang, Y Sun, Z Liu, J Chen, Z Lin; (V) Data analysis and interpretation: J Jiang, M Zhu, L Wang, H Fu, E Long, Z Li, Y Sun, X Wu, D Wang, J Chen, Z Lin, X Liu, H Lin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Haotian Lin. Stata Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. Email: gddlht@aliyun.com; Xiyang Liu. School of Computer Science and Technology, Xidian University, Xi'an 710071, China. Email: xyliu@xidian.edu.cn.

Background: Lens opacity seriously affects the visual development of infants. Slit-illumination images play an irreplaceable role in lens opacity detection; however, these images exhibited varied phenotypes with severe heterogeneity and complexity, particularly among pediatric cataracts. Therefore, it is urgently needed to explore an effective computer-aided method to automatically diagnose heterogeneous lens opacity and to provide appropriate treatment recommendations in a timely manner.

Methods: We integrated three different deep learning networks and a cost-sensitive method into an ensemble learning architecture, and then proposed an effective model called CCNN-Ensemble [ensemble of cost-sensitive convolutional neural networks (CNNs)] for automatic lens opacity detection. A total of 470 slit-illumination images of pediatric cataracts were used for training and comparison between the CCNN-Ensemble model and conventional methods. Finally, we used two external datasets (132 independent test images and 79 Internet-based images) to further evaluate the model's generalizability and effectiveness.

Results: Experimental results and comparative analyses demonstrated that the proposed method was superior to conventional approaches and provided clinically meaningful performance in terms of three grading indices of lens opacity: area (specificity and sensitivity; 92.00% and 92.31%), density (93.85% and 91.43%) and opacity location (95.25% and 89.29%). Furthermore, the comparable performance on the independent testing dataset and the internet-based images verified the effectiveness and generalizability of the model. Finally, we developed and implemented a website-based automatic diagnosis software for pediatric cataract grading diagnosis in ophthalmology clinics.

Conclusions: The CCNN-Ensemble method demonstrates higher specificity and sensitivity than conventional methods on multi-source datasets. This study provides a practical strategy for heterogeneous lens opacity diagnosis and has the potential to be applied to the analysis of other medical images.

Keywords: Cost-sensitive; deep convolutional neural networks (CNNs); ensemble learning; heterogeneous slit-illumination images; pediatric cataract

Submitted Sep 27, 2020. Accepted for publication Jan 12, 2021.
doi: 10.21037/atm-20-6635
View this article at: http://dx.doi.org/10.21037/atm-20-6635

Introduction

Optical imaging technologies play a vital role in the clinical diagnosis and treatment of ophthalmology (1,2). Computational vision approaches for automatic diagnosis of lens opacity have greatly improved the efficiency of ophthalmologists and the entire treatment chain, providing real benefits for patients (3-6). In our previous studies, we applied artificial intelligence methods to detect cataract and then graded lens opacity based on diffuse-light ocular images (7-9). However, the lens opacity grading is solely based on diffuse-light images, which may not be precise as the lens is a three-dimensional object (10-12). The common slit-illumination image offers another effective diagnosis medium and provides an essential supplement to these diffuse-light images (13,14). Therefore, the development of computer vision techniques for slit-illumination images will move the automatic diagnosis of ophthalmic diseases towards a more comprehensive and intelligent strategy.

At present, the existing computer-aided diagnosis methods generally focus on senile cataracts using slit-illumination images (3-5,15). Thresholding localization and support vector regression methods were used to grade the nuclear cataract (16). Recursive convolutional neural networks (CNNs) and support vector regression methods were implemented to enable automatic learning of features for evaluating the severity of nuclear cataracts (17). However, the phenotypes of senile cataracts are relatively simple and fairly homogeneous. The study of such senile cataracts alone will not be sufficient for the development of a computer-aided diagnosis system for lens opacity in complex clinical scenarios. Practical clinical applications need the ability to diagnose heterogeneous lens opacities with high recognition rates (18-20). It is therefore essential to develop an efficient, feasible, and automatic diagnostic system to address heterogeneous slit-illumination images.

The pediatric cataract is a typical lens opacity disease that suffers from severe heterogeneity and complex phenotypes (21-23). Large-scale slit-illumination images of pediatric cataracts were collected from the long-term Childhood Cataract Program of the Chinese Ministry of Health (CCPMOH) project (24), which covered a wide variety of lens opacities. In addition, the imbalance between the categories is an inevitable problem in pediatric cataract diagnosis (21,25), where the number of positive samples is relatively smaller than the number of negative samples. This can easily cause the classifiers to produce a higher false-negative rate. Therefore, these datasets represent an ideal medium for the exploration of the appropriate computational vision methods required to adapt to complex clinical application scenarios.

Recently, 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. To develop an effective and efficient computer vision method for analysis of these heterogeneous slit-illumination images, we integrated three deep CNNs with different structures (AlexNet, GoogLeNet and ResNet50) (26-28) and a cost-sensitive algorithm (33,34) into an ensemble learning framework and created the CCNN-Ensemble model (ensemble of cost-sensitive CNNs). The three CNNs with their different structures were used to improve both the overall recognition rate and stability of the model. The cost-sensitive algorithm was used to address the imbalanced dataset problem and thus significantly reduce the model's false-negative rate. We performed detailed experiments to compare the performance of the CCNN-Ensemble method with that of conventional methods in three grading indices of lens opacity. We also used two external datasets (an independent testing dataset and an Internet-based dataset) to validate the method's versatility and stability. Finally, potential computer-aided diagnostic software was developed and deployed for use by ophthalmologists and their patients in clinical applications.

We present the following article in accordance with the STARD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-6635).

Methods

Dataset

The slit-illumination datasets consist of the following three parts: the training and validation dataset, the independent

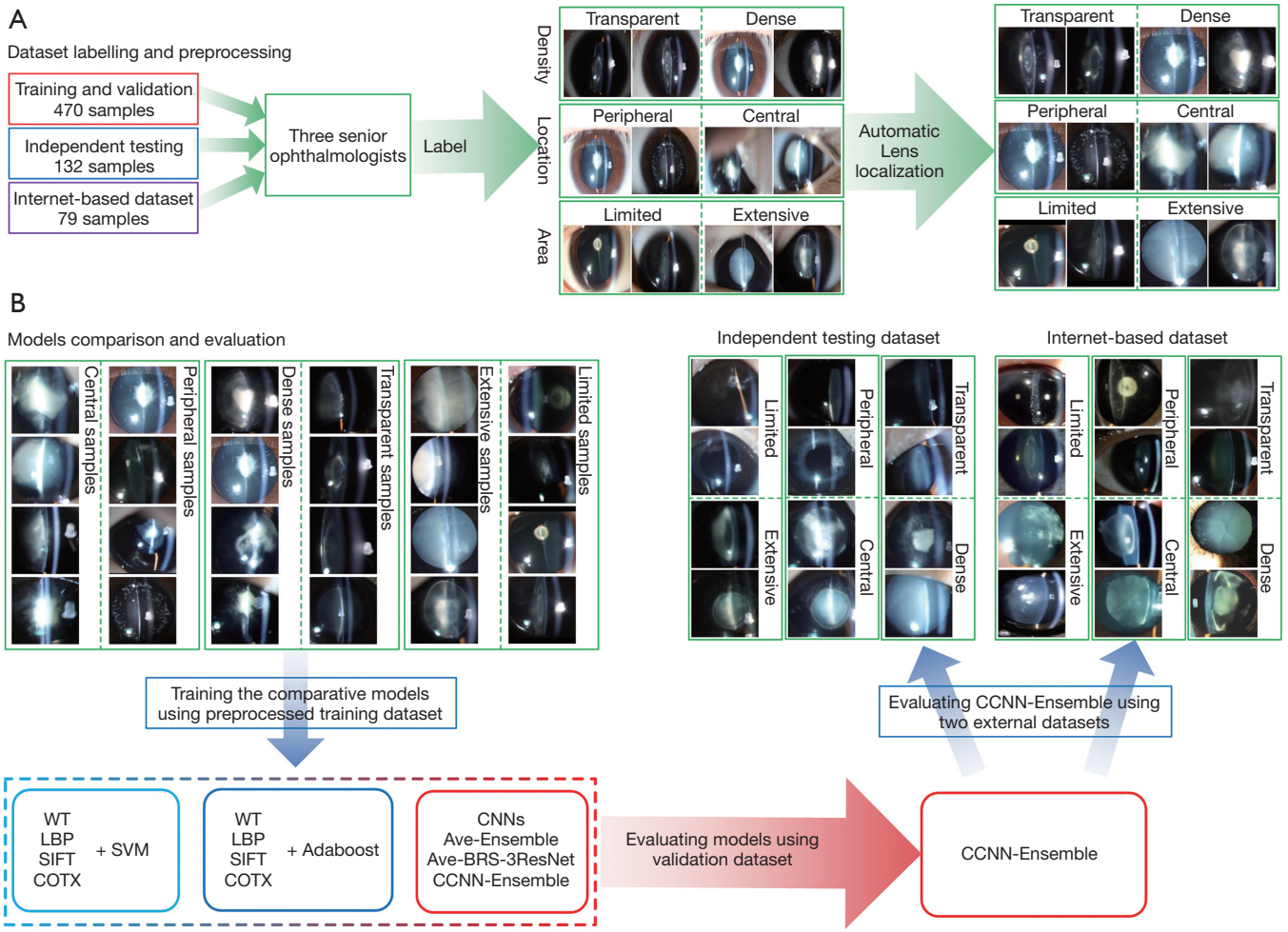

Figure 1 Dataset preparation and performance evaluation of multiple methods. (A) Dataset labelling and preprocessing. Four hundred and seventy training and validation samples and 132 independent test samples were derived from samples provided by the Zhongshan Ophthalmic Center of Sun Yat-sen University; 79 Internet-based samples were collected using the Baidu and Google search engines. Each image was independently graded and labeled by three senior ophthalmologists; subsequently, the images were cropped automatically using twice-applied Canny detection and Hough transformation. (B) Model comparison and evaluation. The training and validation dataset was used to train and evaluate the performances of the different methods and select the best model. Independent testing and Internet-based datasets were also used to evaluate the stability and generalizability of the CCNN-Ensemble method. WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Adaboost, adaptive boosting ensemble learning; Ave-Ensemble, ensemble learning of three different CNNs (AlexNet, GoogLeNet, and ResNet50) with an averaging technique; Ave-BRS-3ResNet, ensemble learning of three ResNet50 architectures with batch random selection and averaging techniques; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

testing dataset, and the Internet-based dataset. A total of 470 training and validation datasets were derived from the routine examinations between June 2015 and February 2020 at Zhongshan Ophthalmic Center of Sun Yat-sen University (*Figure 1A*) (24). 132 independent testing images were selected randomly in advance from the Zhongshan Ophthalmic Center; 79 Internet-based images were collected using a keyword search (including words such as congenital cataract, infant, and pediatric) of the Baidu and Google search engines. In total, there were 470 individuals in the training and validation datasets and 132 individuals in the independent testing dataset. All individuals underwent

Table 1 Distributions of slit-illumination datasets in terms of three grading indices

Datasets	Total number	Opacity area		Opacity density		Opacity location	
		Limited	Extensive	Transparent	Dense	Peripheral	Central
Training and validation datasets	470	275	195	260	210	274	196
Independent testing dataset	132	91	41	104	28	100	32
Internet-based dataset	79	19	60	18	61	16	63

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). The age of the subjects in training, validation, and independent testing datasets is 18.96±10.61 months (mean ± SD). The study was approved by the Ethics Committee of Zhongshan Ophthalmic Center of Sun Yat-sen University (NO.: 2017KYPJ096) and adhered to the tenets of the Declaration of Helsinki (as revised in 2013). Written informed consent was obtained from all the study participants' parents or legal guardians.

There are no special pixel requirements for the enrolled images provided that the lens area of the image is retained. To ensure grade labeling accuracy, three senior ophthalmologists jointly determine the grade of each image and comprehensively evaluate its severities in terms of three lens lesion indices (opacity area, density, and location) (7,9). An opacity area that covers more than half of the pupil is defined as extensive; otherwise, it is defined as limited. An opacity density that completely blocks the light is labelled as dense; otherwise, it is defined as transparent. An opacity location that fully covers the visual axis of the pupil is called central; otherwise, it is called peripheral. The collected datasets covered a variety of pediatric cataracts, which were divided into limited and extensive categories for the area, dense and transparent categories for density, and central and peripheral categories for location, as shown in *Table 1*.

Preprocessing and model evaluation

We preprocessed all labeled datasets using twice-applied Canny detection and Hough transformation (35,36) to acquire the lens region of interest and eliminate surrounding noise zones such as the eyelids and the sclera (*Figure 1A*). The detailed procedures and methods of automatic lens cropping are consistent with our previous research (7,9). The localized images were subsequently resized to a size of 256×256 pixels and were then input into the computational vision models. Using these training and validation datasets, we performed a five-fold cross-validation procedure to compare and evaluate the performances of the different models (*Figure 1B*). Four representative handcrafted features (WT: wavelet transformation; LBP: local binary pattern; SIFT: scale-invariant feature transform; and COTE: color and texture features) (8,9,37-39) were selected and combined with support vector machine (SVM) and adaptive boosting (Adaboost) classifiers for performance comparison. In addition, three single-classifier CNNs (AlexNet, GoogLeNet, and ResNet50) and two common ensemble learning methods (Ave-Ensemble and Ave-BRS-3ResNet) based on CNNs were performed to compare with CCNN-Ensemble. 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. After the selection of the optimal CCNN-Ensemble model, we further verified its effectiveness and stability using the two external datasets (the independent testing dataset and the Internet-based dataset).

Statistical analysis

To provide a full assessment of the superiority of the CCNN-Ensemble method when compared with the conventional methods, we calculated several evaluation metrics, including accuracy, sensitivity, specificity, F1-measure, and G-mean, as follows.

$$Accuracy = (TP + TN)/(TP + FN + TN + FP) \quad [1]$$

$$Sensitivity(Recall) = TP/(TP + FN) \quad [2]$$

$$Specificity = TN/(TN + FP) \quad [3]$$

$$Precision = TP/(TP + FP) \quad [4]$$

$$F1\text{-}measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad [5]$$

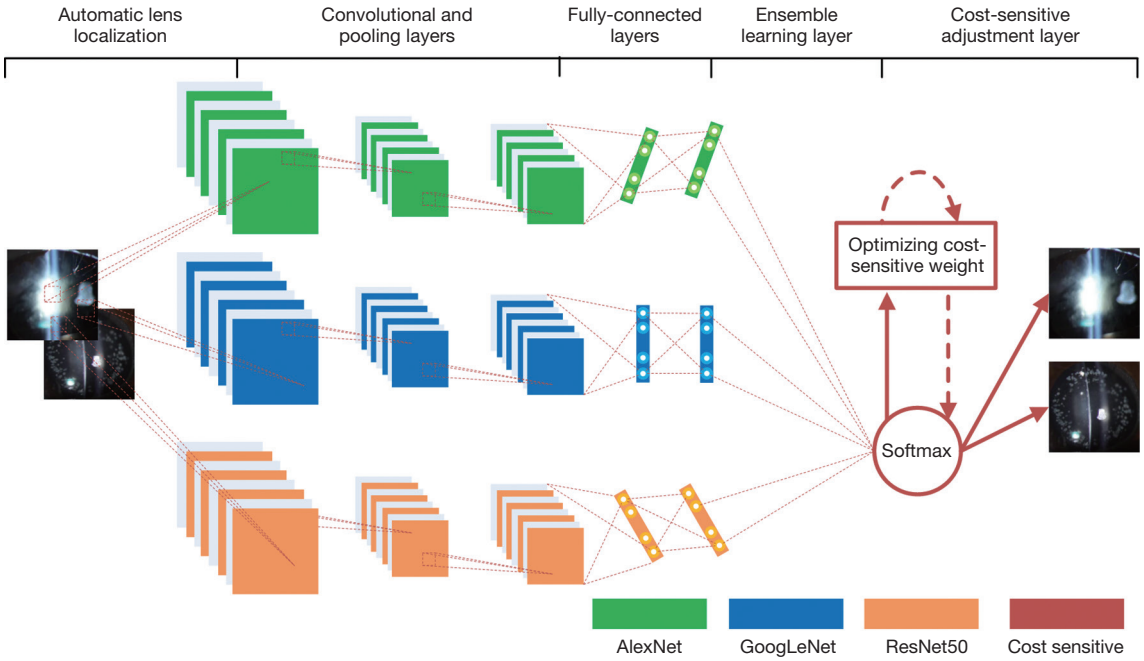

Figure 2 Framework of the CCNN-Ensemble method. The preprocessed images were input into three parallel deep learning CNNs (AlexNet, GoogLeNet, and ResNet50) with different network structures for feature extraction and classification; a unified ensemble learning of CNNs was then used to improve the recognition rate of the classifier. The cost-sensitive layer was used to adjust the costs of the positive and negative samples in the loss function to address the imbalanced dataset problem. CNN, convolutional neural network; AlexNet, eight-layer Alex CNN; GoogLeNet, 22-layer inception CNN developed by Google researchers; ResNet50, 50-layer residual CNN.

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}} \quad [6]$$

where *TP*, *FP*, *TN*, and *FN* denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. Accuracy, sensitivity, and specificity are the most commonly used evaluation measures. The F1-measure and G-mean (40) indicators simultaneously consider the accuracies of both classes and can thus effectively measure the recognition abilities of models in the case of an imbalanced dataset. Additionally, three more vital objective measures—the receiver operating characteristic curve (ROC), the area under the ROC curve (AUC), and the precision recall curve (PR)—were used for visual comparison and analysis. Five-fold cross-validation was applied to calculate the mean and standard deviation of the above evaluation metrics. All statistical analyses were conducted using python 3.7.8.

Overall framework of CCNN-Ensemble

As shown in *Figure 2*, the overall diagnosis framework of the CCNN-Ensemble consists primarily of three deep CNN models (GoogLeNet, AlexNet, and ResNet50), a cost-sensitive adjustment layer, ensemble learning, dataset augmentation technology, and transfer learning. The three heterogeneous CNN models, as classifiers, were employed to construct the ensemble learning framework to enhance the recognition rates of the algorithms. The cost-sensitive adjustment layer was used to manage the imbalanced dataset problem, and the dataset augmentation and transfer learning processes were adopted to overcome the overfitting problem and accelerate model convergence. The technical details are described below.

Ensemble learning of multiple heterogeneous CNNs

We used three heterogeneous CNNs (AlexNet, GoogLeNet and ResNet50) to form the ensemble learning framework (*Figure 2*). The AlexNet CNN, which was proposed by Krizhevsky (26), performed image classification and won first prize in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, mainly used convolutional layers, overlapping pooling, nonsaturating rectified linear

units (ReLUs) and three fully-connected layers to construct an eight-layer CNN. Subsequently, a number of variants CNNs were proposed to enhance its recognition rate and incorporated many emerging technologies. In particular, a 22-layer inception deep network was achieved by Google researchers (27) that were based on the Hebbian principle, an intuition of multi-scale processing, filter aggregation, average pooling, and auxiliary classifier technologies. Kaiming He then used the residual connection scheme, batch normalization, and scale operations to establish a 50-layer ultra-deep residual CNN (ResNet50) (28). Because the above CNNs implemented different principles and techniques, their network structures show distinct heterogeneity, and this can effectively improve the recognition rate of the ensemble learning model.

In order to adequately utilize the advantages of the three CNNs, we implemented a two-stage ensemble learning scheme. Specifically, in the first stage, starting with the initial parameters of models pre-trained on the ImageNet dataset, three CNNs with different structures were trained using transfer learning, respectively. Thus, the optimal parameters of each CNN were obtained. In the second stage, the Softmax functions of the three CNNs were removed, the high-level features of the CNNs were merged into the same cost-sensitive Softmax classification function to construct a unified ensemble CNN. The learning rate of the feature extraction layers was set to one-tenth of the ensemble learning layer. The transfer learning method was adopted to fully train the ensemble learning layer and fine-tune the previous feature extraction layers. Through the above two-stage ensemble learning scheme, three different types of CNNs can complement their shortcomings, which is beneficial to improve the overall performance of intelligent diagnosis for pediatric cataract.

Transfer learning

Because the number of medical images is very small, the fully-trained deep learning system cannot adequately optimize the millions of trainable parameters from scratch and this can easily lead to overfitting. Transfer learning (41,42) is a critical technology for application to such small datasets that allows the model to be trained from a better starting point and uses the color, texture, and shape characteristics that have been learned from natural images. Fine-tuning allowed the final trained CNN model to obtain the unique features of the ophthalmic images and also overcame the overfitting problem. Additionally, data augmentation methods, including transformed images and horizontal reflections (26,43), were adopted to accelerate the convergence of the models.

Cost-sensitive method and optimization process

To address the imbalanced dataset problem of the slit-illumination images effectively, the cost-sensitive approach (33,34,44) was adopted to adjust the cost-sensitive weight of the positive samples in the loss function (*Figure 2*). Specifically, we discriminatively determined the cost of misclassification of the different classes and assigned a larger cost-sensitive weight to the positive class. For one iterative training stage, n samples were selected at random to form a training dataset $\{[x^{(1)}, y^{(1)}], [x^{(2)}, y^{(2)}],...,[x^{(n)}, y^{(n)}]\}$, where $x^{(i)} \in R^l$ and $y^{(i)} \in \{1,...,k\}$. Here, $x^{(i)}$ denotes the features of the i-th sample and $y^{(i)}$ is the category label. The cost-sensitive loss function can be expressed as shown in Eq. [7].

$$F(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} I\left\{y^{(i)} = j\right\} * CS\left\{y^{(i)} = \text{positive class}\right\} * \log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{s=1}^{k} e^{\theta_s^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_{ij}^2 \quad [7]$$

where n, m, k and θ denote the number of training samples, the number of input neurons, the number of classes, and trainable parameters, respectively. $I\{y^{(i)} = j\}$ represents the indicator function ($I\{y^{(i)}$ *is equal to* $j\} = 1$ and $I\{y^{(i)}$ *is not equal to* $j\} = 0$) while $CS\{y^{(i)} = \text{positive class}\}$ is the cost-sensitive weight function ($CS\{y^{(i)}$ *is the positive class lable*$\} = C$ and $CS\{y^{(i)}$ *is the negative class lable*$\} = 1$). Using a grid-search procedure, we determined that the value of the effective cost-sensitive weight parameter C was within the interval [4–6]. $\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_{ij}^2$ is a weight decay term that is applied to penalize the larger trainable weights. To obtain the optimal trainable weights θ^* (see Eq. [8]), we needed to minimize $F(\theta)$ using mini-batch gradient descent (Mini-batch-GD) (45) as shown in Eq. [9].

$$\theta^* = \arg\min_{\theta} F(\theta) \quad [8]$$

$$\nabla_{\theta_j} F(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big[PW\left\{y^{(i)} = \text{positive class}\right\} * x^{(i)} * (I\{y^{(i)} = j\} - p(y^{(i)} = j \mid x^{(i)}; \theta))\Big] + \lambda\theta_j \quad [9]$$

Visualization heatmaps

To verify the reasonability and effectiveness of the CCNN-Ensemble, the Gradient-weighted Class Activation Mapping (Grad-CAM) (46) visualization technique was employed to generate the heatmaps for highlighting the disease-related regions on which the diagnosis model focused most. The Grad-CAM is an explainable technique for CNN-based models, which utilized the gradients of any target concept flowing into last convolutional layer to produce a localization map highlighting remarkable regions in an image for predicting the concept.

Experimental environment

In this study, we implemented dataset preprocessing, automatic lens region of interest (ROI) localization, conventional feature extraction, the SVM and Adaboost classifiers, and uniform dataset partitioning for cross-validation using MATLAB R2014a (8,9). The training, validation, and testing procedures of three single-classifier CNNs (AlexNet, GoogleNet, and ResNet50) and three ensemble learning methods were all performed in parallel using eight Nvidia Titan X graphics processing units (GPUs) based on the Caffe toolbox (47) in the Ubuntu 16.4 OS. 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 and ensemble learning methods. The initial learning rate was set at 0.001 and successively reduced by one tenth of the original value after every 500 iterations; a total of 2,000 iterations were performed. We set the mini-batch size to 32 on one GPU and used eight GPUs; we thus acquired a total of 256 samples in every iteration and calculated the average value of these samples to update the trainable parameters. Appropriate settings for these parameters can ensure better performance and rapid convergence for the CCNN-Ensemble method.

Results

To achieve an effective solution to assist in the diagnosis of pediatric cataracts using slit-illumination images, we explored five different models, including four conventional features, four Adaboost ensemble methods, three single-classifier CNNs, two conventional ensemble learning based on CNNs, and the CCNN-Ensemble method. First, we trained and compared the performances of these methods on the training and validation datasets to obtain the optimal CCNN-Ensemble method. Then, we used two external datasets to provide further evaluation of the robustness and the clinical effectiveness of the CCNN-Ensemble. Finally, we developed and deployed cloud-based software to serve patients that were located in remote areas.

Performance comparison of CCNN-Ensemble with conventional features and Adaboost ensemble methods

After application of the five-fold cross-validation (48), we compared the performances of the nine intelligent algorithms for the lens opacity in terms of the three grading indices (opacity area, density, and location). We calculated three main indicators—accuracy (ACC), specificity (SPE), and sensitivity (SEN) (*Figure 3*)— along with more detailed test results with means and standard deviations (*Table 2* and Tables S1,S2). First, when using the conventional feature methods, both the ACC and SEN indicators are low; for example, the SEN of the LBP method is less than 70% for all grading indices. Second, after the application of the Adaboost ensemble learning methods, the SEN indicator is greatly improved, whereas the value of the SPE indicator is reduced. As a result, the ACC is almost equal to the performance of the conventional feature methods (*Figure 3*). Notably, the SEN of the SIFT method increased from 76.41% to 84.62%, whereas the SPE decreased from 76.73% to 65.45% for opacity area grading (*Figure 3* and *Table 2*); the SEN of the LBP method increased from 68.88% to 81.10%, whereas the SPE again decreased from 80.27% to 73.34% for opacity location grading (*Figure 3* and Table S2). The comparison results for the other feature methods and the Adaboost ensemble learning methods are also similar. Third, the CCNN-Ensemble method provided significantly improved recognition rates for all grading indices (*Figure 3*). All the average ACCs were maintained at 92% or more, while both the SPE and the SEN were satisfactory for the grading opacity area (92.00% and 92.31%), the opacity density (93.85% and 91.43%), and the opacity location (95.25% and 89.29%). Similarly, the F1-measure, G-mean, and AUC indicators also showed values of more than 90% (*Table 2* and Tables S1,S2).

Additionally, we used the ROC and PR curves to compare the performances of the above methods (*Figure 4*, Figures S1,S2). The ROC curve of the CCNN-Ensemble is close to the upper-left area of the graph and the PR

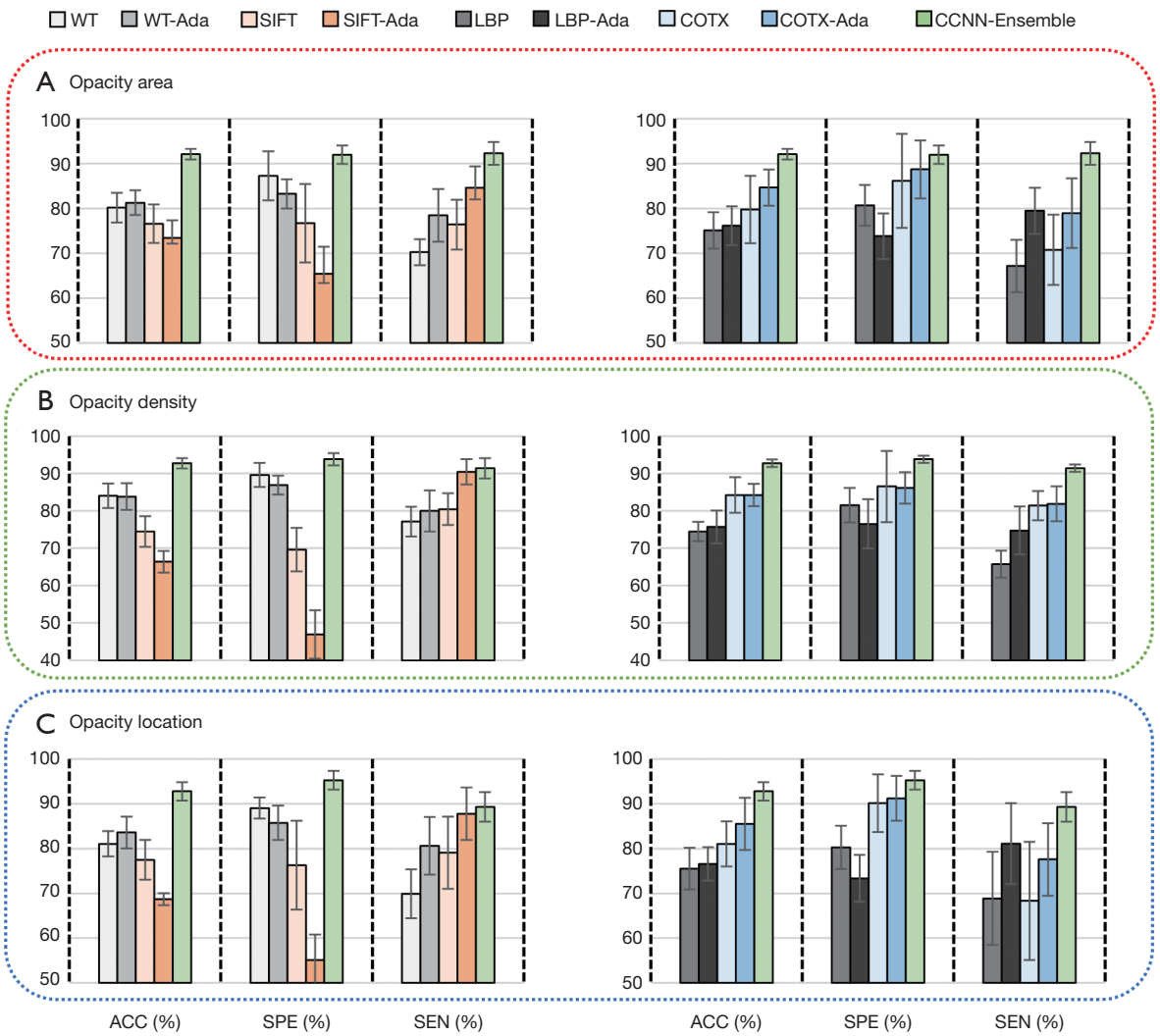

Figure 3 Performance comparisons of the different methods for the three grading indices. Performance comparisons of conventional features, Adaboost ensemble learning, and CCNN-Ensemble methods for the lens opacity area, opacity density, and opacity location, respectively. The sensitivity of Adaboost ensemble learning methods is greatly improved over the conventional feature methods, whereas their specificity indicator is reduced and the accuracy has no significant improvement. The CCNN-Ensemble method outperforms other conventional features and Adaboost ensemble approaches and offers exceptional accuracy, specificity, and sensitivity in terms of three grading indices of lens opacity: area (92.13%, 92.00%, and 92.31%), density (92.77%, 93.85%, and 91.43%) and location (92.76%, 95.25%, and 89.29%). ACC, accuracy; SPE, specificity; SEN, sensitivity; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Ada, adaptive boosting ensemble learning; WT-Ada, adaptive boosting ensemble learning with wavelet transformation feature; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

curve shows a similar performance. All the AUC indicators were maintained at more than 0.969 for the three grading indices. This indicates that the CCNN-Ensemble method is superior to conventional features and Adaboost ensemble learning methods.

Performance comparison of CCNN-Ensemble with single-classifier CNNs and conventional ensemble learning based on CNNs

To further verify the superiority of the CCNN-Ensemble method, we conducted comparative experiments including

Table 2 Performance comparison of CCNN-Ensemble with conventional features and Adaboost ensemble methods in opacity area grading

Method	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC (%)
WT	80.21 (3.33)[§]	87.27 (5.45)	70.26 (2.92)	74.73 (3.50)	78.26 (2.86)	87.47 (2.87)
WT-Adaboost	81.28 (2.77)	83.27 (3.25)	78.46 (5.90)	77.61 (3.59)	80.76 (3.13)	89.68 (2.54)
LBP	75.11 (4.09)	80.73 (4.56)	67.18 (5.85)	69.11 (5.09)	73.59 (4.26)	83.45 (3.82)
LBP- Adaboost	76.17 (4.36)	73.82 (5.08)	79.49 (5.13)	73.48 (4.69)	76.56 (4.36)	83.69 (3.38)
SIFT	76.60 (4.32)	76.73 (8.76)	76.41 (5.56)	73.12 (3.56)	76.35 (3.90)	85.66 (4.05)
SIFT- Adaboost	73.40 (3.98)	65.45 (6.03)	84.62 (4.80)	72.56 (3.67)	74.33 (3.94)	85.61 (4.15)
COTX	79.79 (7.52)	86.18 (10.5)	70.77 (7.82)	74.62 (8.54)	77.93 (7.02)	87.22 (5.22)
COTX- Adaboost	84.68 (4.02)	88.73 (6.48)	78.97 (7.78)	81.01 (4.92)	83.53 (4.34)	91.07 (2.85)
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)*

[§], mean (standard deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Adaboost, adaptive boosting ensemble learning; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

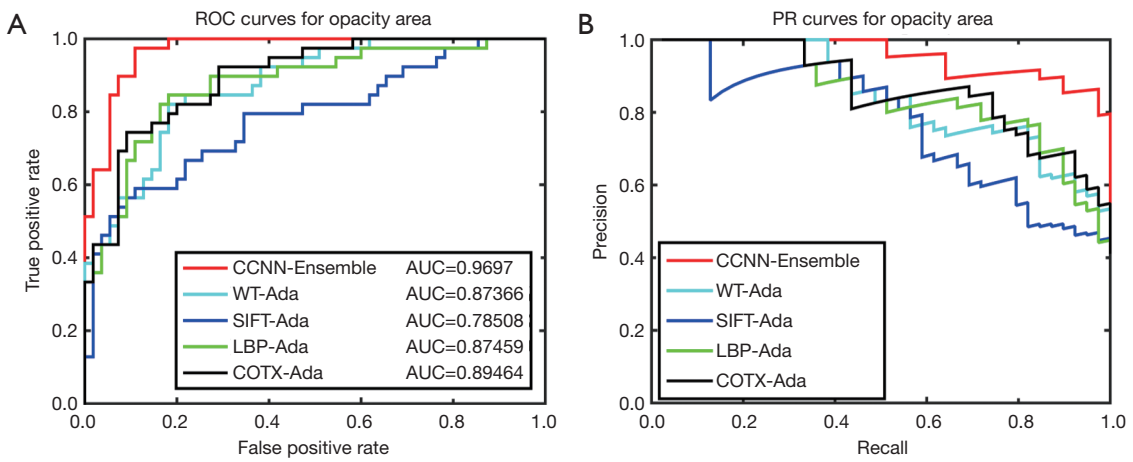

Figure 4 ROC and PR curves for the different methods in opacity area grading. (A) ROC curves and AUC values for the CCNN-Ensemble method and four comparison methods: WT-Ada, SIFT-Ada, LBP-Ada, and COTE-Ada. (B) PR curves for the CCNN-Ensemble method and the four comparison methods. WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Ada, adaptive boosting ensemble learning; WT-Ada, adaptive boosting ensemble learning with wavelet transformation feature; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks; ROC, receiver operating characteristic curve; AUC, area under the ROC curve; PR, precision recall curve.

three ensemble learning methods (Ave-Ensemble, Ave-BRS-3ResNet, and CCNN-Ensemble) and three single-classifier CNNs (AlexNet, GoogLeNet, and ResNet50). Detailed results of three grading indices of lens opacity (opacity area, density, and localization) were shown in *Table 3* and Supplementary Tables S3,S4. From the results of opacity area grading, we had 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

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)*

[§], mean (standard deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; Ave-Ensemble, ensemble learning of three different CNNs (AlexNet, GoogLeNet and ResNet50) with an averaging technique; Ave-BRS-3ResNet, ensemble learning of three ResNet50 architectures with batch random selection and averaging techniques; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks;

Table 4 Quantitative evaluation of the CCNN-Ensemble method using two external datasets

External Datasets	Grading	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC (%)
Independent testing dataset	Opacity area	94.70	96.70	90.24	91.36	93.42	96.94
	Opacity density	93.18	94.23	89.29	84.75	91.72	97.70
	Opacity location	93.18	94.00	90.63	86.57	92.30	98.13
Internet-based dataset	Opacity area	89.87	89.47	90.00	93.10	89.74	94.65
	Opacity density	88.61	88.89	88.52	92.31	88.71	95.63
	Opacity location	87.34	87.50	87.30	91.67	87.40	93.06

ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

the 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 (Tables S3,S4).

Performance in independent testing dataset

To ensure an adequate investigation of the generalizability and the effectiveness of the CCNN-Ensemble method, we used an independent testing dataset for further validation of the proposed method. A total of 132 slit-illumination images were selected randomly in advance from the Zhongshan Ophthalmic Center (details are given in the Methods section). Using the expert group's decisions for reference, we presented detailed quantitative evaluation results (as shown in *Table 4*) and performance comparison (*Figure 5A*). We also reported the ROC and PR curves for the three grading indices: opacity area, density, and location (*Figure 5A*). The experimental results indicated that the performance of the CCNN-Ensemble method on the independent testing dataset is almost equal to that of the validation dataset, with the ACC and the SPE being maintained at more than 93% and 94%, respectively, and the SEN values are 90.24%, 89.29% and 90.63% for the opacity grading area, density, and location, respectively.

Performance in Internet-based dataset

In addition, we also collected 79 slit-illumination images from the Internet (details are given in the Methods section). While the quality of these images varied significantly, the CCNN-Ensemble was still able to detect the appropriate cases with a higher recognition rate. In the same manner, we obtained detailed prediction results (given in *Table 4*), intuitive

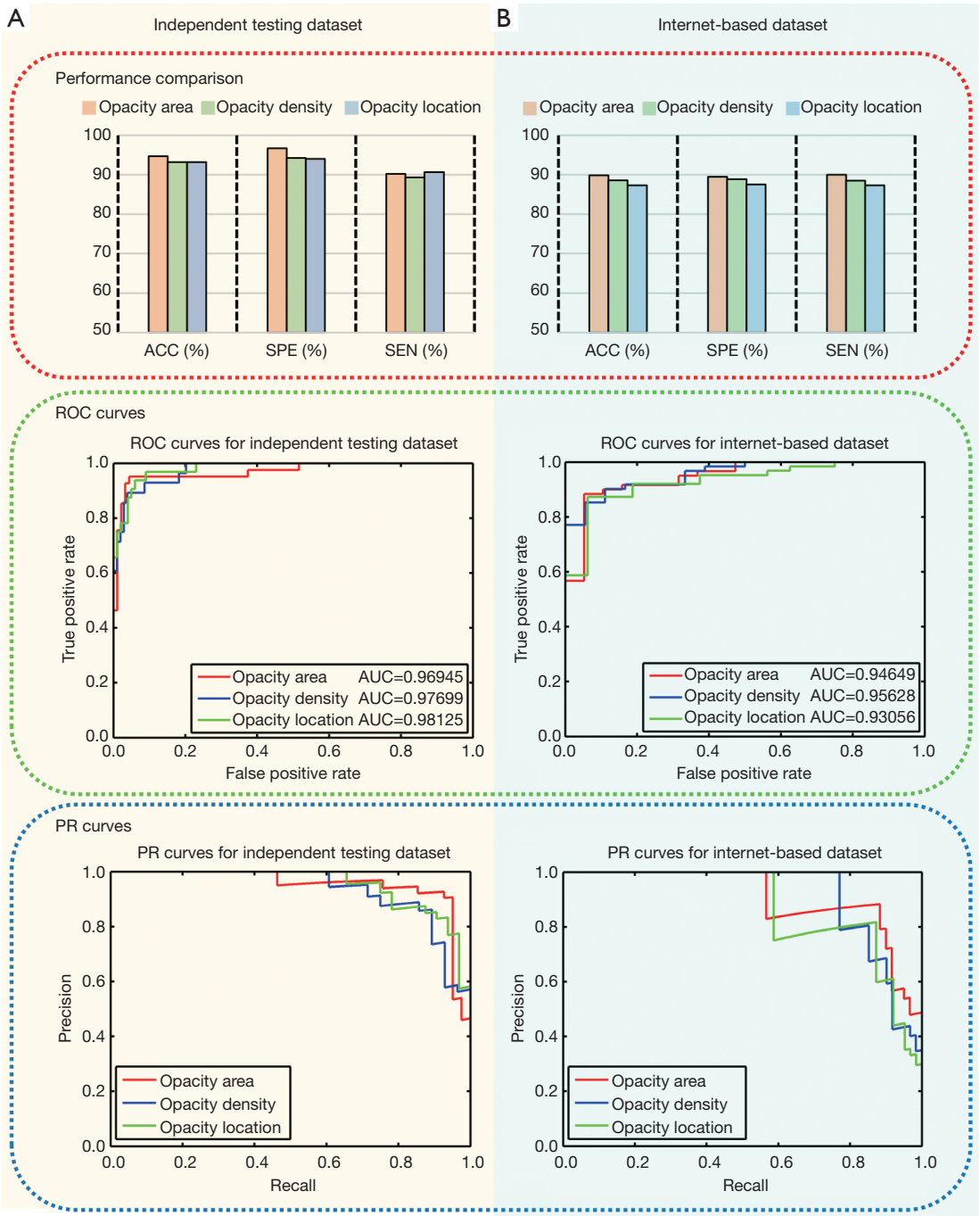

Figure 5 Performance analysis results for the CCNN-Ensemble on two external datasets. (A) The performance comparison, ROC curves, and PR curves of the CCNN-Ensemble method for lens opacity area, density, and location grading on the independent testing dataset. (B) The performance comparison, ROC curves, and PR curves for lens opacity area, density, and location grading on Internet-based dataset. The model performances are satisfactory when applied to the two external datasets, independent test images: area (94.70%, 96.70%, and 90.24%), density (93.18%, 94.23%, and 89.29%) and location (93.18%, 94.00%, and 90.63%); internet-based images: area (89.87%, 89.47%, and 90.00%), density (88.61%, 88.89%, and 88.52%) and location (87.34%, 87.50%, and 87.30%), indicating that the model is universal and effective. ACC, accuracy; SPE, specificity; SEN, sensitivity; ROC, receiver operating characteristic curve; AUC, area under the ROC curve; PR, precision recall curve.

comparison graphs for the main indicators (ACC, SPE, and SEN), the ROC curve, and the PR curve (*Figure 5B*). Specifically, the CCNN-Ensemble method also offered satisfactory accuracy, specificity, and sensitivity in terms of opacity area (89.87%, 89.47%, and 90.00%), opacity density (88.61%, 88.89%, and 88.52%), and opacity location (87.34%, 87.50%, and 87.30%), respectively.

Interpretability analysis of CCNN-Ensemble for opacity area grading

Using the Grad-CAM technique, three heatmaps were obtained simultaneously via the CCNN-Ensemble, which were associated with the Alexnet, GoogLeNet, and ResNet50, respectively. In the independent testing dataset, four representative slit-illumination images of opacity area grading and their heatmaps are displayed in *Figure 6*. The highlighted colors in the heatmap indicate the opacity areas on which the network was based to make a decision.

Web-based software

To serve both patients and ophthalmologists located in remote areas, we developed and deployed an automatic diagnosis software based on cloud service (http://www.cc-cruiser.com:5007/SignIn), which included user registration, an image upload module, a prediction module, regular re-examinations, sample downloads, and instructions. For evaluation and trial, we provided a test user (ws) and its password (ws) of the diagnosis software. Before using the website for diagnosis, the users needed to submit personal information including age, gender, and telephone number to complete the registration process. This registration process allowed the doctor to contact patients who were diagnosed with serious conditions, and also prevented the illegal use of our software. After registration, either the patient or the ophthalmologist can upload the slit-illumination images for diagnosis; the software can then perform image preprocessing, make three grading predictions, and provide a final treatment recommendation. Our software can diagnose multiple images simultaneously. A total of 30 sample images were available for download, and our e-mail address and telephone number were also provided for all registered patients.

Discussion

The inferior performance of conventional feature methods when applied to diagnosis using the slit-illumination images is mainly attributed to the following two causes. First, the conventional feature methods use handcrafted descriptors to represent the original images, which are completely reliant on the designer's experience and operator techniques, and which cannot learn statistical features from the existing large dataset. Second, the conventional feature methods and the SVM classifier do not take the problem of the imbalanced dataset into account, and this results in the final predictions being biased towards the majority class and ignoring the minority class (i.e., the positive samples). Therefore, these methods lead to inferior overall accuracy and lower sensitivity.

The Adaboost ensemble learning methods led to moderate improvement of the recognition rates when compared with the conventional feature methods because they train and apply multiple classifiers jointly to determine the final grading results. Simultaneously, an under-sampling method is incorporated into Adaboost to address the imbalanced dataset. Therefore, the sensitivity of the methods is greatly enhanced, but this improvement leads to the reduction of the specificity. The overall accuracy rate is almost equal to that obtained when using the conventional feature methods alone.

The CCNN-Ensemble method is significantly superior to the above methods in terms of all grading indices, which was attributed to the following four improvements. First, the CCNN-Ensemble method does not need to design any feature descriptor manually because it learns high-level and statistical features directly from the original images. Second, we use three different CNNs for ensemble learning, so that they can learn the different characteristics from three different perspectives to enable joint determination of the final prediction. This ensemble of multiple CNN technologies is beneficial in enhancing the overall performance. Third, the cost-sensitive approach is integrated into the CCNN-Ensemble method and takes greater account of the minority class to ensure that the sensitivity indicator is valid for the imbalanced dataset. In addition, transfer learning is applied to our model to enable fine-tuning of the trainable parameters from a better starting point, thus making it easier to jump out from the local minimum. As a result, the higher accuracy and specificity performances are maintained while the sensitivity is also greatly enhanced.

The performance of three ensemble learning methods was superior to those of the three single-classifier CNNs. The reason is that multiple classifiers in the ensemble

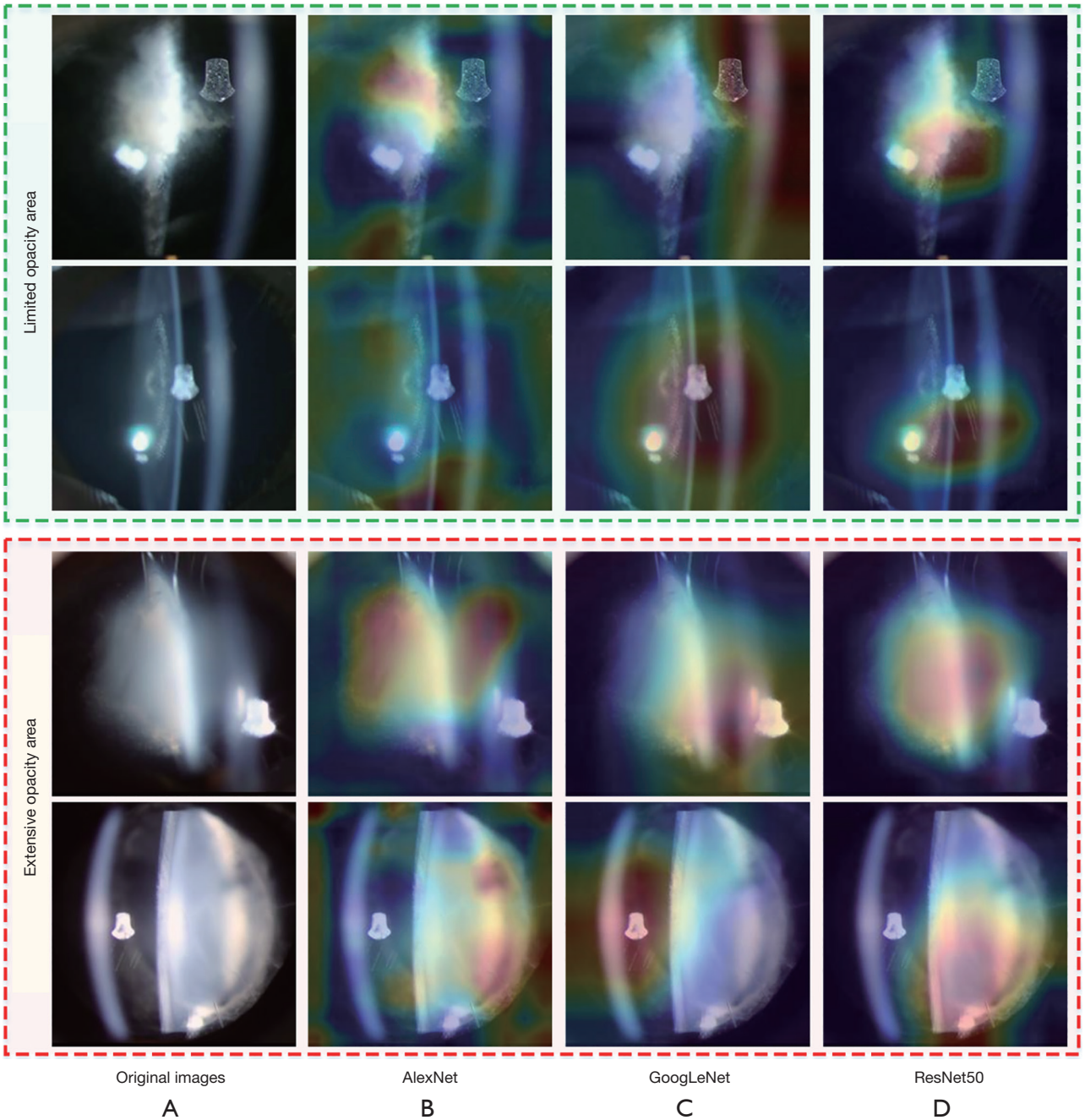

Figure 6 The representative heatmaps of CCNN-Ensemble in opacity area grading using Grad-CAM. (A) The original slit-illumination images. (B,C,D) The visualization heatmaps generated from Alexnet, GoogLeNet, and ResNet50 in the CCNN-Ensemble method. The upper two rows indicate negative samples with limited opacity area, and the lower two rows represent positive samples with extensive opacity area. Grad-CAM, Gradient-weighted Class Activation Mapping.

learning methods are complementary to each other and their advantages are fully utilized, thereby improving the performance of a single classifier. Compared with the conventional ensemble learning methods, the CCNN-Ensemble method combines three heterogeneous CNNs, and performs a two-stage transfer learning to fully optimize

the parameters of the three networks, thereby further enhancing the performance of ensemble learning. In addition, by analyzing the heatmap shown in *Figure 6*, it finds that all these three classifiers can capture the location of lens opacity, although the highlighted areas are slightly different. The Grad-CAM technique further corroborates the effectiveness of the proposed diagnosis system. Interpretability analysis of heatmap provides strong evidence for the acceptance of our CCNN-Ensemble in ophthalmic clinics.

The CCNN-Ensemble method also demonstrated better performance on two external datasets, and their recognition rates were almost equal to that of the validation dataset. This indicates that the proposed approach is insensitive to different data sources, and its generalizability and robustness are better than those of the conventional methods. These experimental conclusions provide sufficient evidence to justify the application of the CCNN-Ensemble method in complex clinical scenarios.

Based on our proposed method, automated diagnostic software was developed and deployed to serve patients and ophthalmologists remotely in the form of a cloud service, which provided important clinical value for pediatric cataract diagnosis. By accessing our automatic diagnostic software remotely, any patient can upload slit-illumination images and can then quickly obtain prediction results and an appropriate treatment recommendation. Therefore, this remotely-aided diagnosis method avoided doctors from performing tedious examinations and helped patients located in remote areas. In addition, this work can also provide a teaching role for junior doctors.

However, several limitations of this study should be mentioned. First, multiple CNNs with different structures are integrated into the architecture. Although the strategy of ensemble learning significantly improves the accuracy, it is slightly less cost-effective due to the high requirement of the computing resource than a single CNN model. Second, our model is solely based on the slit-illumination image, which is insufficient to identify the lens opacity in occasional situations. Combining the electronic medical records and other optical images may provide valuable supplements for the comprehensive assessment of lens opacity. Third, the robustness and stability of our method are required to be verified before the further generalization of other medical situations. Despite the above limitations, this study provides a practical strategy for heterogeneous lens opacity diagnosis with promising performance validated in multi-source datasets. Further studies with the integration of electronic medical records and more optical images will pave the way for the wide-range clinical application of our work.

Conclusions

In this paper, we proposed a feasible and automated CCNN-Ensemble method for the effective diagnosis of pediatric cataracts using heterogeneous slit-illumination images. We integrated three deep CNNs and cost-sensitive technology to construct an ensemble learning method that could identify the severity of lens opacity based on three grading indices. The experimental results and comparison analyses verified that the proposed method is superior to other conventional methods. The performance of the CCNN-Ensemble method on two external datasets indicated its improved robustness and generalizability. Finally, a set of cloud-based automatic diagnostic software was produced for use by both patients and ophthalmologists. This research could provide a helpful reference for the analysis of other medical images and will help to promote the use of artificial intelligence techniques in clinical applications.

Acknowledgments

Funding: This study was funded by the National Key R&D Program of China (2018YFC0116500), the National Natural Science Foundation of China (81770967), the National Natural Science Fund for Distinguished Young Scholars (81822010), the Science and Technology Planning Projects of Guangdong Province (2018B010109008, 2019B030316012), and the Fundamental Research Funds for the Central Universities (JBX180704). The sponsor or funding organization had no role in the design or conduct of this research.

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-6635

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-6635

Peer Review File: Available at http://dx.doi.org/10.21037/atm-20-6635

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.

org/10.21037/atm-20-6635). 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. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of Zhongshan Ophthalmic Center of Sun Yat-sen University (No.: 2017KYPJ096) and written informed consent was obtained from all the study participants' parents or legal guardians.

References

1. Bernardes R, Serranho P, Lobo C. Digital ocular fundus imaging: a review. Ophthalmologica 2011;226:161-81.
2. Ng EY, Acharya UR, Suri JS, et al. Image Analysis and Modeling in Ophthalmology. CRC Press, 2014.
3. Zhang Z, Srivastava R, Liu H, et al. A survey on computer aided diagnosis for ocular diseases. BMC Med Inform Decis Mak 2014;14:80.
4. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-75.
5. Armstrong GW, Lorch AC. A (eye): A Review of Current Applications of Artificial Intelligence and Machine Learning in Ophthalmology. Int Ophthalmol Clin 2020;60:57-71.
6. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Exp Ophthalmol 2019;47:128-39.
7. 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.
8. Wang L, Zhang K, Liu X, et al. Comparative analysis of image classification methods for automatic diagnosis of ophthalmic images. Sci Rep 2017;7:41545.
9. Liu X, Jiang J, Zhang K, et al. Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. PloS One 2017;12:e0168606.
10. Klein BEK, Klein R, Linton KLP, et al. Assessment of cataracts from photographs in the Beaver Dam Eye Study. Ophthalmology 1990;97:1428-33.
11. Reid JE, Eaton E. Artificial intelligence for pediatric ophthalmology. Curr Opin Ophthalmol 2019;30:337-46.
12. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine 2019;9:52-9.
13. Chylack LT Jr, Wolfe JK, Singer DM, et al. The Lens Opacities Classification System III. The Longitudinal Study of Cataract Study Group. Arch Ophthalmol 1993;111:831-6.
14. Kumar S, Yogesan K, Constable I. Telemedical diagnosis of anterior segment eye diseases: validation of digital slit-lamp still images. Eye 2009;23:652-60.
15. Kolhe S, Guru MSK. Cataract Classification and Grading: A Survey. Int J Innov Res Comput Commun Eng 2015;3:10749-55.
16. Li H, Lim JH, Liu J, et al. A Computer-Aided Diagnosis System of Nuclear Cataract. IEEE Trans Biomed Eng 2010;57:1690-8.
17. Gao X, Lin S, Wong TY. Automatic feature learning to grade nuclear cataracts based on deep learning. IEEE Trans Biomed Eng 2015;62:2693-701.
18. Amaya L, Taylor D, Russell-Eggitt I, et al. The morphology and natural history of childhood cataracts. Surv Ophthalmol 2003;48:125-44.
19. Wu X, Long E, Lin H, et al. Prevalence and epidemiological characteristics of congenital cataract: a systematic review and meta-analysis. Sci Rep 2016;6:28564.
20. Medsinge A, Nischal KK. Pediatric cataract: challenges and future directions. Clin Ophthalmol 2015;9:77.
21. Lenhart PD, Courtright P, Wilson ME, et al. Global challenges in the management of congenital cataract: proceedings of the 4th International Congenital Cataract Symposium held on March 7, 2014, New York, New York. J AAPOS 2015;19:e1-8.
22. Ellis FJ. Management of pediatric cataract and lens opacities. Curr Opin Ophthalmol 2002;13:33-7.
23. Wilson ME, Trivedi RH, Pandey SK. Pediatric cataract surgery: techniques, complications, and management.

Lippincott Williams & Wilkins, 2005.
24. Lin H, Long E, Chen W, et al. Documenting rare disease data in China. Science 2015;349:1064.
25. Chen W, Long E, Chen J, et al. Timing and approaches in congenital cataract surgery: a randomised controlled trial. Lancet 2016;388:S52.
26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Adv Neural Inf Process Syst 2012;25:1097-105.
27. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: 2015;1-9.
28. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. arXiv preprint arXiv:151203385 2015.
29. Kumar A, Kim J, Lyndon D, et al. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomed Health Inform 2017;21:31-40.
30. Rasti R, Teshnehlab M, Phung SL. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. Pattern Recognition 2017;72:381-90.
31. Bermejo-Peláez D, Ash SY, Washko GR, et al. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. Sci Rep 2020;10:338.
32. Chen H, Dou Q, Wang X, et al., editors. Mitosis detection in breast cancer histology images via deep cascaded networks. Thirtieth AAAI conference on artificial intelligence; 2016.
33. Ali S, Majid A, Javed SG, et al. Can-CSC-GBE: Developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data. Comput Biol Med 2016;73:38-46.
34. Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 2006;18:63-77.
35. Daugman J. New methods in iris recognition. IEEE Trans Syst Man Cybern B Cybern 2007;37:1167-75.
36. Masek L. Recognition of human iris patterns for biometric identification. The University of Western Australia, 2003.
37. Yang JJ, Li J, Shen R, et al. Exploiting ensemble learning for automatic cataract detection and grading. Comput Methods Programs Biomed 2016;124:45-57.
38. Guo L, Yang J-J, Peng L, et al. A computer-aided healthcare system for cataract classification and grading based on fundus image analysis. Comput Ind 2015;69:72-80.
39. Huang W, Chan KL, Li H, et al. A Computer Assisted Method for Nuclear Cataract Grading From Slit-Lamp Images Using Ranking. IEEE Trans Med Imaging 2011;30:94-107.
40. Tang Y, Zhang YQ, Chawla NV, et al. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B Cybern 2009;39:281-8.
41. Shin HC, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016;35:1285-98.
42. Ravishankar H, Sudhakar P, Venkataramani R, et al. Understanding the Mechanisms of Deep Transfer Learning for Medical Images. arXiv preprint arXiv:170406040.
43. Cireşan D, Meier U, Masci J, et al. Multi-column deep neural network for traffic sign classification. Neural Netw 2012;32:333-8.
44. Krawczyk B, Schaefer G, Woźniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artif Intell Med 2015;65:219-27.
45. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Lechevallier Y, Saporta G. editors. Proceedings of COMPSTAT'2010. Physica-Verlag HD. doi: 10.1007/978-3-7908-2604-3_16.
46. Selvaraju RR, Cogswell M, Das A, et al. editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017.
47. Jia Y, Shelhamer E, Donahue J, et al. 'Caffe: convolutional architecture for fast feature embedding', arXiv preprint arXiv:14085093, 2014.
48. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence 1995;14:1137-45.

Cite this article as: Jiang J, Wang L, Fu H, Long E, Sun Y, Li R, Li Z, Zhu M, Liu Z, Chen J, Lin Z, Wu X, Wang D, Liu X, Lin H. Automatic classification of heterogeneous slit-illumination images using an ensemble of cost-sensitive convolutional neural networks. Ann Transl Med 2021;9(7):550. doi: 10.21037/atm-20-6635

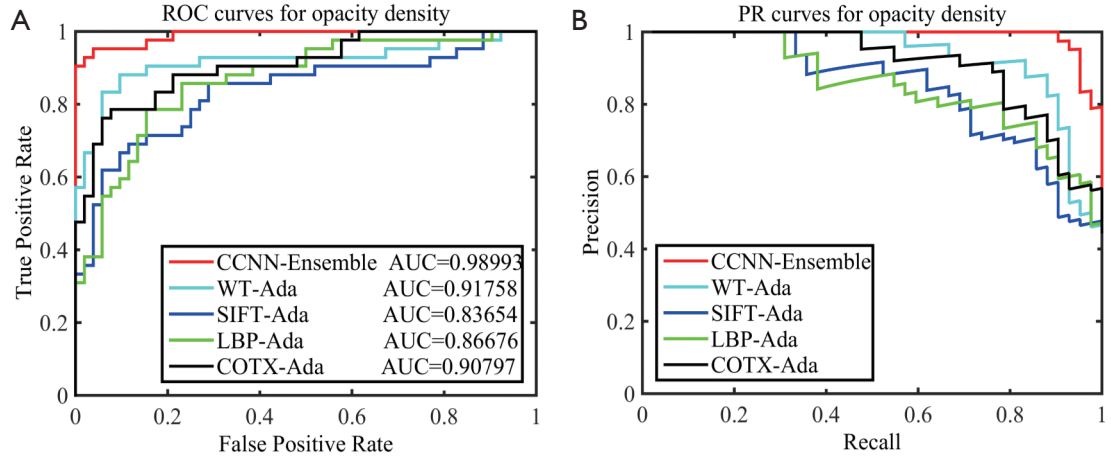

Figure S1 ROC and PR curves for the different methods in opacity density grading. (A) ROC curves and AUC values for the CCNN-Ensemble method and four comparison methods: WT-Ada, SIFT-Ada, LBP-Ada, and COTE-Ada. (B) PR curves for the CCNN-Ensemble method and the four comparative methods. ROC, receiver operating characteristic curve; AUC, area under the ROC curve; PR, precision-recall curve; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks; Ada, adaptive boosting ensemble learning; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features.

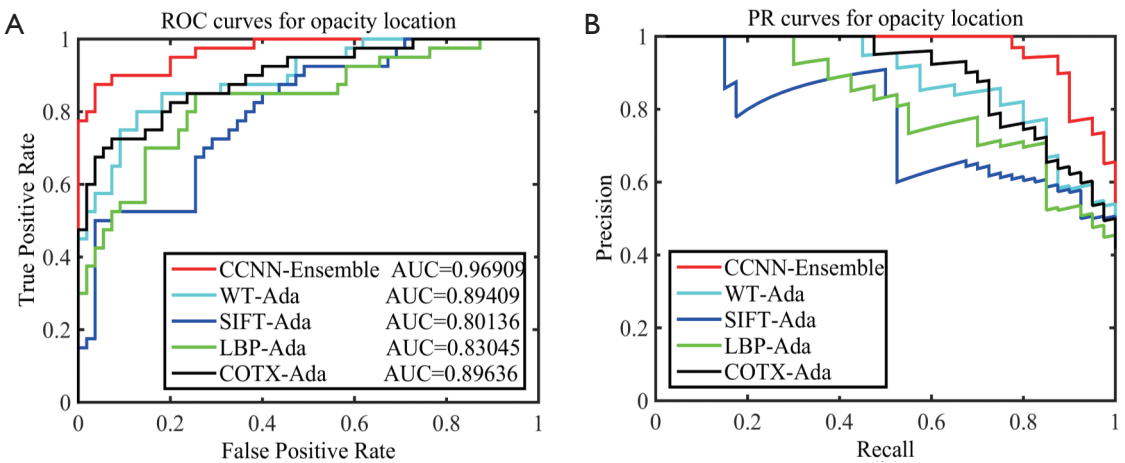

Figure S2 ROC and PR curves for the different methods in opacity location grading. (A) ROC curves and AUC values for the CCNN-Ensemble method and four comparison methods: WT-Ada, SIFT-Ada, LBP-Ada, and COTE-Ada. (B) PR curves for the CCNN-Ensemble method and the four comparison methods. ROC, receiver operating characteristic curve; AUC, area under the ROC curve; PR, precision-recall curve; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks; Ada, adaptive boosting ensemble learning; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; WT-Ada, adaptive boosting ensemble learning with wavelet transformation feature.

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

Table S1 Performance comparison of CCNN-Ensemble with conventional features and Adaboost ensemble methods in the opacity density grading

Method	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC (%)
WT	84.04 (3.28) [§]	89.62 (3.22)	77.14 (3.98)	81.20 (3.83)	83.13 (3.36)	90.05 (1.69)
WT- AdaBoost	83.83 (3.56)	86.92 (2.51)	80.00 (5.48)	81.50 (4.31)	83.37 (3.79)	90.33 (1.92)
LBP	74.47 (2.61)	81.54 (4.63)	65.71 (3.61)	69.70 (2.78)	73.14 (2.46)	81.38 (4.09)
LBP- AdaBoost	75.74 (4.41)	76.54 (6.58)	74.76 (6.43)	73.35 (4.54)	75.52 (4.34)	82.30 (4.17)
SIFT	74.47 (4.12)	69.62 (5.83)	80.48 (4.26)	73.83 (3.93)	74.79 (4.07)	83.73 (1.96)
SIFT- AdaBoost	66.38 (2.87)	46.92 (6.46)	90.48 (3.37)	70.66 (1.79)	64.99 (3.67)	83.91 (1.92)
COTX	84.26 (4.72)	86.54 (9.52)	81.43 (3.91)	82.38 (4.25)	83.78 (4.25)	89.78 (4.50)
COTX- AdaBoost	84.26 (2.95)	86.15 (4.17)	81.90 (4.64)	82.29 (3.31)	83.95 (3.01)	92.06 (2.74)
CCNN-Ensemble	*92.77 (1.39)*	*93.85 (1.61)*	*91.43 (2.71)*	*91.86 (1.62)*	*92.62 (1.47)*	*98.01 (0.85)*

[§], Mean (Standard Deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Adaboost, adaptive boosting ensemble learning; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

Table S2 Performance comparison of CCNN-Ensemble with conventional features and Adaboost ensemble methods in the opacity location grading

Method	ACC (%)	SPE (%)	SEN (%)	F_M (%)	G_M (%)	AUC (%)
WT	81.06 (2.81) [§]	89.04 (2.34)	69.91 (5.46)	75.42 (3.85)	78.85 (3.29)	89.34 (2.69)
WT- AdaBoost	83.60 (3.53)	85.75 (3.82)	80.62 (6.42)	80.35 (4.38)	83.08 (3.79)	90.56 (3.41)
LBP	75.52 (4.67)	80.27 (4.85)	68.88 (10.4)	69.90 (6.35)	74.15 (5.35)	81.70 (5.64)
LBP- AdaBoost	76.58 (3.73)	73.34 (5.21)	81.10 (9.03)	74.16 (4.56)	76.94 (3.92)	82.81 (6.18)
SIFT	77.47 (4.44)	76.30 (9.90)	79.05 (8.06)	74.56 (4.09)	77.32 (4.14)	85.46 (3.47)
SIFT- AdaBoost	68.72 (1.31)	55.11 (5.66)	87.76 (5.84)	70.03 (1.14)	69.34 (1.83)	85.08 (3.40)
COTX	81.05 (5.04)	90.12 (6.46)	68.33 (13.2)	74.52 (8.88)	78.00 (7.37)	90.18 (3.58)
COTX- AdaBoost	85.52 (5.79)	91.21 (5.00)	77.58 (8.07)	81.68 (7.32)	84.07 (6.16)	91.62 (3.48)
CCNN-Ensemble	*92.76 (2.06)*	*95.25 (2.08)*	*89.29 (3.30)*	*91.14 (2.53)*	*92.21 (2.19)*	*97.29 (1.36)*

[§], Mean (Standard Deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; WT, wavelet transformation; LBP, local binary pattern; SIFT, scale-invariant feature transform; COTE, color and texture features; Adaboost, adaptive boosting ensemble learning; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

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

Table S3 Performance comparison of CCNN-Ensemble with single-classifier CNNs and conventional ensemble learning methods based on CNNs in opacity density grading

Method	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC (%)
AlexNet	88.09 (2.65)[§]	88.85 (3.16)	87.14 (4.64)	86.34 (3.71)	87.95 (2.76)	93.16 (2.32)
GoogLeNet	88.94 (2.68)	89.62 (2.92)	88.10 (4.45)	87.65 (2.73)	88.80 (2.58)	94.55 (2.04)
ResNet50	89.57 (2.75)	90.38 (2.36)	88.57 (4.88)	88.32 (2.28)	89.44 (2.95)	95.46 (1.92)
Ave-Ensemble	90.43 (1.86)	91.15 (2.01)	89.26 (2.80)	89.26 (2.12)	90.29 (2.42)	96.41 (1.64)
Ave-BRS-3ResNet	90.00 (2.33)	90.77 (2.24)	89.05 (3.26)	88.80 (2.30)	89.86 (2.56)	96.23 (1.83)
CCNN-Ensemble	*92.77 (1.39)*	*93.85 (1.61)*	*91.43 (2.71)*	*91.86 (1.62)*	*92.62 (1.47)*	*98.01 (0.85)*

[§], Mean (Standard Deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; Ave-Ensemble, ensemble learning of three different CNNs (AlexNet, GoogLeNet and ResNet50) with an averaging technique; Ave-BRS-3ResNet, ensemble learning of three ResNet50 architectures with batch random selection and averaging techniques; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

Table S4 Performance comparison of CCNN-Ensemble with single-classifier CNNs and conventional ensemble learning methods based on CNNs in opacity location grading

Method	ACC (%)	SPE (%)	SEN (%)	F1_M (%)	G_M (%)	AUC (%)
AlexNet	88.30 (2.78)	90.88 (4.37)	84.71 (3.55)	85.81 (3.05)	87.71 (2.57)	90.71 (3.23)
GoogLeNet	88.72 (2.54)	91.25 (3.28)	85.21 (3.27)	86.30 (3.72)	88.15 (2.38)	92.24 (3.02)
ResNet50	89.58 (2.02)	91.61 (3.12)	86.76 (2.89)	87.43 (2.29)	89.11 (2.01)	93.70 (2.63)
Ave-Ensemble	90.64 (1.45)	92.70 (1.98)	87.77 (2.01)	88.66 (1.14)	90.19 (1.24)	94.83 (1.64)
Ave-BRS-3ResNet	90.21 (1.68)	92.34 (2.21)	87.26 (2.47)	88.14 (1.34)	89.74 (1.52)	94.05 (1.95)
CCNN-Ensemble	*92.76 (2.06)*	*95.25 (2.08)*	*89.29 (3.30)*	*91.14 (2.53)*	*92.21 (2.19)*	*97.29 (1.36)*

[§], Mean (Standard Deviation). ACC, accuracy; SPE, specificity; SEN, sensitivity; F1_M, F1-measure; G_M, G-mean; AUC, area under the receiver operating characteristic curve; Ave-Ensemble, ensemble learning of three different CNNs (AlexNet, GoogLeNet and ResNet50) with averaging technique; Ave-BRS-3ResNet, ensemble learning of three ResNet50 architectures with batch random selection and averaging techniques; CCNN-Ensemble, ensemble learning of cost-sensitive convolutional neural networks.

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