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 Electronics Engineering, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China;

[2]School of Computer Science and Technology, Xidian University, Xi'an, 710071, China;

[3]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, 510060, China;

[4]School of Mathematics and Statistics, Xidian University, Xi'an, 710071, China;

[#]These authors contributed equally to this work.

[*]**Co-corresponding authors:**

Email: haot.lin@hotmail.com (H.T.L) and xyliu@xidian.edu.cn (X.Y.L)

Abstract

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 ECCNN-Ensemble (ensemble of cost-sensitive convolutional neural networks) 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 therefore provides a practical strategy for heterogeneous lens opacity diagnosis and has the potential to be applied to the analysis of other medical images.

Key words: cost-sensitive; deep convolutional neural networks; ensemble learning; heterogeneous slit-illumination images; pediatric cataract.

Background

Optical imaging technologies play a vital role in 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 the classification of diffuse-light ocular images [7-9]. However, diagnosis that is solely dependent on diffuse-light images will inevitably miss a substantial proportion of potential ophthalmology patients [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, 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 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, 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 exploration of the appropriate computational vision methods required to adapt to complex clinical application scenarios.

To develop an effective and efficient computer vision method for analysis of these heterogeneous slit-illumination images, we integrated three deep convolutional neural networks (AlexNet, GoogLeNet and ResNet) [26-28] and a cost-sensitive algorithm [29, 30] into an ensemble learning framework and created the CCNN-Ensemble model (ensemble of cost-sensitive convolutional neural networks). The three convolutional neural networks (CNNs) with their different structures were used to improve both 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 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.

Methods

Dataset

The slit-illumination datasets consist of the following three parts: the training and validation dataset, the independent testing dataset, and the Internet-based dataset. A total of 470 training and validation datasets were derived from the routine examinations of the Zhongshan Ophthalmic Center in Sun Yat-sen University (Fig. 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.

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 area, dense and transparent categories for density, and central and peripheral categories for location, as shown in Table 1.

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 datset	470	275	195	260	210	274	196
Independent testing datset	132	91	41	104	28	100	32
Internet-based datset	79	19	60	18	61	16	63

Preprocessing and model evaluation

We preprocessed all labeled datasets using twice-applied Canny detection and Hough transformation [31, 32] to acquire the lens region of interest and eliminate surrounding noise zones such as the eyelids and the sclera (Fig. 1a). 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 (Fig. 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, 33-35] were selected and combined with support vector machine (SVM) and adaptive boosting (Adaboost) classifiers for performance comparision. After 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).

Evaluation metrics

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.

$$Accuracy = (TP + TN)/(TP + FN + TN + FP) \tag{1}$$

$$Sensitivity(Recall) = TP/(TP + FN) \tag{2}$$

$$Specificity = TN/(TN + FP) \tag{3}$$

$$Precision = TP/(TP + FP) \tag{4}$$

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

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

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 [36] 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.

Overall framework of CCNN-Ensemble

As shown in Fig. 2, the overall diagnosis framework of the CCNN-Ensemble consists primarily of three deep CNN models (GoogLeNet, AlexNet and ResNet), 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 CNNs

We used three heterogeneous CNNs (AlexNet, GoogLeNet and ResNet) to form the ensemble learning framework (Fig. 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 of CNN method 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 was based on the Hebbian principle, 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 (ResNet) [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 convolutional neural networks, 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 [37, 38] 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, 39], 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 [29, 30, 40] was adopted to adjust the cost-sensitive weight of the positive samples in the loss function (Fig. 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.

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

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. $I\left\{y^{(i)} = j\right\}$ represents the indicator function ($I\left\{y^{(i)}\ is\ equal\ to\ j\right\} = 1$ and $I\left\{y^{(i)}\ is\ not\ equal\ to\ j\right\} = 0$) while $CS\left\{y^{(i)} = positive\ class\right\}$ is the cost-sensitive weight function ($CS\left\{y^{(i)}\ is\ the\ positive\ class\ lable\right\} = C$ and $CS\left\{y^{(i)}\ is\ the\ negative\ class\ label\right\} = 1$). 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 a mini-batch gradient descent (Mini-batch-GD) [41] as shown in Eq. 9.

$$\theta^* = \arg\min_{\theta} F(\theta) \tag{8}$$

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

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 CCNN-Ensemble training, validation and testing procedures were all performed in parallel using eight Nvidia Titan X graphics processing units (GPUs) based on the Caffe toolbox [42] in the Ubuntu 16.4 OS. 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 2000 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 three different methods, including four conventional features, four Adaboost ensemble methods, 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 of CCNN-Ensemble and conventional methods

After application of the five-fold cross-validation [43], we compared the performances of the nine intelligent algorithms for diagnosis of 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) (Fig. 3) – along with more detailed test results with means and standard deviations (Table 2 and Supplementary Table 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 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 (see Fig. 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 (Fig. 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 (Fig. 3 and Supplementary 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 (Fig. 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 Supplementary Table S1-S2).

Table 2. Performance comparison of the different 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)***

Notes: 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; [§]Mean (Standard Deviation).

Additionally, we used the ROC and PR curves to compare the performances of the above methods (Fig. 4, Supplementary Fig. S1-S2). The ROC curve of the CCNN-Ensemble is close to the upper-left area of the graph and the PR 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 the conventional features and Adaboost ensemble learning methods.

Performance in independent testing dataset

Table 3. 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

Notes: ACC: accuracy; SPE: specificity; SEN: sensitivity; F1_M: F1-measure; G_M: G-mean; AUC: area under the receiver operating characteristic curve.

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 3) and performance comparisons of the ACC, SPE and SEN indicators (Fig. 5a). We also reported the ROC and PR curves for the three grading indices: opacity area, density and location (Fig. 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 3), intuitive comparison graphs for the main indicators (ACC, SPE and SEN), the ROC curve, and the PR curve (Fig. 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.

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. 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 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 sensivity 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 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 the different data sources, and that 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 freed the 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 an 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 depended 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 wide-range clinical application of our work.

Conclusions

In this paper, we proposed a feasible and automated CCNN-Ensemble method for 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 analysis of other medical images and will help to promote the use of artificial intelligence techniques in clinical applications.

Supplementary file

Supplementary file 1: Detailed performance comparison of the different methods in terms of the opacity density and location grading. The detailed comparison results of the different methods in terms of the opacity density and location grading are presented in Supplementary files.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Zhongshan Ophthalmic Center of Sun Yat-sen University. Written informed consent was obtained from all the study participants' parents or legal guardian.

Consent for publication

Not applicable.

Availability of data and materials

The datasets of the current study are available from the corresponding author on reasonable request.

Abbreviations

CNN: convolutional neural network; CCNN-Ensemble: ensemble of cost-sensitive convolutional neural networks; ResNet: residual convolutional neural network; Adaboost: adaptive boosting ensemble learning; SVM: support vector machine; LBP: local binary pattern; WT: wavelet transformation; SIFT: scale-invariant feature transform; COTE: color and texture features; ReLUs: rectified linear units; Mini-batch-GD: mini-batch gradient descent; ACC: accuracy; SPE: specificity; SEN: sensitivity; F1_M: F1-measure; G_M: G-mean; ROC: receiver operating characteristic curve; AUC: area under the ROC curve; PR: precision recall curve; ROI: region of interest; CCPMOH: Childhood Cataract Program of the Chinese Ministry of Health.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

J.W.J., H.T.L. and X.Y.L designed the research; J.W.J., E.P.L, L.M.W., Z.W.L., M.M.Z. and R.Y.L conducted the study; E.P.L, Z.Z.L., Z.L.L., J.J.C and D.N.W. collected the data; J.W.J., Y.B.S., H.R.F. and L.M.W. were responsible for coding and computer-aided diagnosis software; M.M.Z. supported the mathematical theory; J.W.J., X.H.W., Z.W.L., J.J.C., H.R.F. and Y.B.S analyzed and completed the experimental results; and J.W.J., E.P.L., H.T.L. and X.Y.L. co-wrote the manuscript. All the authors discussed the results and reviewed the manuscript.

Acknowledgements

We am grateful thanks to Dr. Lin Liu, Shuai Wang, and Fan Liu for their helpful guidance and suggestion with this project.

Funding

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

Authors' information

Correspondence and requests for materials should be addressed to H.T.L. (haot.lin@hotmail.com) or X.Y.L. (xyliu@xidian.edu.cn).

References

1. Bernardes R, Serranho P, Lobo C. Digital ocular fundus imaging: a review. Ophthalmologica. 2011;226(4):161-81.
2. Ng EY, Acharya UR, Suri JS, Campilho A. Image Analysis and Modeling in Ophthalmology: CRC Press; 2014.
3. Zhang Z, Srivastava R, Liu H, Chen X, Duan L, Kee Wong DW, 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, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology. 2019;103(2):167-75.
5. Armstrong GW, Lorch AC. A (eye): A review of current applications of artificial intelligence and machine learning in ophthalmology. International Ophthalmology Clinics. 2020;60(1):57-71.
6. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clinical & experimental ophthalmology. 2019;47(1):128-39.
7. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering. 2017;1:0024.
8. Wang L, Zhang K, Liu X, Long E, Jiang J, An Y, et al. Comparative analysis of image classification methods for automatic diagnosis of ophthalmic images. Scientific Reports. 2017;7.
9. Liu X, Jiang J, Zhang K, Long E, Cui J, Zhu M, 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(3):e0168606.
10. Klein BEK, Klein R, Linton KLP, Magli YL, Neider MW. Assessment of cataracts from photographs in the Beaver Dam Eye Study. Ophthalmology. 1990;97(11):1428-33.
11. Reid JE, Eaton E. Artificial intelligence for pediatric ophthalmology. Current opinion in ophthalmology. 2019;30(5):337-46.

12. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, 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. Archives of Ophthalmology. 1993;111(6):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(3):652-60.
15. Kolhe S, Guru MSK. Cataract Classification and Grading: A Survey. 2015.
16. Li H, Lim JH, Liu J, Mitchell P, Tan AG, Wang JJ, et al. A Computer-Aided Diagnosis System of Nuclear Cataract. IEEE Transactions on Biomedical Engineering. 2010;57(7):1690-8.
17. Gao X, Lin S, Wong TY. Automatic feature learning to grade nuclear cataracts based on deep learning. IEEE Transactions on Biomedical Engineering. 2015;62(11):2693-701.
18. Amaya L, Taylor D, Russell-Eggitt I, Nischal KK, Lengyel D. The morphology and natural history of childhood cataracts. Survey of ophthalmology. 2003;48(2):125-44.
19. Wu X, Long E, Lin H, Liu Y. Prevalence and epidemiological characteristics of congenital cataract: a systematic review and meta-analysis. Scientific Reports. 2016;6:28564.
20. Medsinge A, Nischal KK. Pediatric cataract: challenges and future directions. Clinical ophthalmology (Auckland, NZ). 2015;9:77.
21. Lenhart PD, Courtright P, Wilson ME, Lewallen S, Taylor DS, Ventura MC, 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. Journal of American Association for Pediatric Ophthalmology and Strabismus. 2015;19(2):e1-e8.
22. Ellis FJ. Management of pediatric cataract and lens opacities. Current opinion in ophthalmology. 2002;13(1):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, Liu Y. Documenting rare disease data in China. Science. 2015;349(6252).
25. Chen W, Long E, Chen J, Liu Z, Lin Z, Cao Q, et al. Timing and approaches in congenital

cataract surgery: a randomised controlled trial. The Lancet. 2016;388:S52.

26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012;25(2):2012.
27. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1-9.
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint arXiv:151203385. 2015.
29. Ali S, Majid A, Javed SG, Sattar M. Can-CSC-GBE: Developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data. Computers in biology and medicine. 2016;73:38-46.
30. Zhou Z-H, Liu X-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering. 2006;18(1):63-77.
31. Daugman J. New methods in iris recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B. 2007;37(5):1167-75.
32. Masek L. Recognition of human iris patterns for biometric identification. The University of Western Australia. 2003;2.
33. Yang J-J, Li J, Shen R, Zeng Y, He J, Bi J, et al. Exploiting ensemble learning for automatic cataract detection and grading. Computer methods and programs in biomedicine. 2016;124:45-57.
34. Guo L, Yang J-J, Peng L, Li J, Liang Q. A computer-aided healthcare system for cataract classification and grading based on fundus image analysis. Computers in Industry. 2015;69:72-80.
35. Huang W, Chan KL, Li H, Lim JH, Liu J, Wong TY. A computer assisted method for nuclear cataract grading from slit-lamp images using ranking. IEEE Transactions on Medical Imaging. 2011;30(1):94-107.
36. Tang Y, Zhang Y-Q, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009;39(1):281-8.

37. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016.
38. Ravishankar H, Sudhakar P, Venkataramani R, Thiruvenkadam S, Annangi P, Babu N, et al. Understanding the mechanisms of deep transfer learning for medical images. arXiv preprint arXiv:170406040. 2017.
39. Ciresan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. Computer Vision and Pattern Recognition (CVPR). 2012:3642-9.
40. Krawczyk B, Schaefer G, Woźniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artificial intelligence in medicine. 2015;65(3):219-27.
41. Bottou L. Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010. 2010:177-86.
42. Jia YaS, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:14085093. 2014.
43. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence. 1995;14(2):1137-45.

Figure legends

Fig. 1. Dataset preparation and performance evaluation of multiple methods. (a) Dataset labelling and preprocessing. 470 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 the generalizability of the CCNN-Ensemble method. Notes: 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.

Fig. 2. Framework of the CCNN-Ensemble method. The preprocessed images were input into three parallel deep learning CNNs (AlexNet, GoogLeNet, and ResNet) 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. Notes: CNN: convolutional neural network; AlexNet: eight-layer Alex CNN; GoogLeNet: 22-layer inception CNN developed by Google researchers; ResNet: 50-layer residual CNN.

Fig. 3. Performance comparisons of the different methods for the three grading indices. Images **(a)–(c)** show 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%). Notes: 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.

Fig. 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. Notes: 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.

Fig. 5. Performance analysis results for the CCNN-Ensemble on two external datasets. (a) The performance comparision, ROC curves and PR curves of the CCNN-Ensemble method for lens opacity area, density and location grading on independent testing dataset. **(b)** The performance comparision, 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. Notes: ACC: accuracy; SPE: specificity; SEN: sensitivity; ROC: receiver operating characteristic curve; AUC: area under the ROC curve; PR: precision recall curve.