



OPEN ACCESS

EDITED BY

Wei Chi,
Sun Yat-sen University, China

REVIEWED BY

Xianjun Zhu,
Sichuan Academy of Medical Sciences
and Sichuan Provincial People's Hospital,
China
Gilbert Yong San Lim,
SingHealth, Singapore

*CORRESPONDENCE

Qinxiang Zheng,
✉ zhengqinxiang@aliyun.com
Jiewei Jiang,
✉ jiangjw924@126.com
Wei Chen,
✉ chenweimd@wmu.edu.cn

†These authors have contributed equally
to this work

RECEIVED 30 March 2023

ACCEPTED 20 July 2023

PUBLISHED 28 July 2023

CITATION

Xie H, Li Z, Wu C, Zhao Y, Lin C, Wang Z,
Wang C, Gu Q, Wang M, Zheng Q, Jiang J
and Chen W (2023), Deep learning for
detecting visually impaired cataracts
using fundus images.
Front. Cell Dev. Biol. 11:1197239.
doi: 10.3389/fcell.2023.1197239

COPYRIGHT

© 2023 Xie, Li, Wu, Zhao, Lin, Wang,
Wang, Gu, Wang, Zheng, Jiang and Chen.
This is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Deep learning for detecting visually impaired cataracts using fundus images

He Xie^{1†}, Zhongwen Li^{2†}, Chengchao Wu³, Yitian Zhao^{2,4},
Chengmin Lin⁵, Zhouqian Wang¹, Chenxi Wang¹, Qinyi Gu¹,
Minye Wang¹, Qinxiang Zheng^{1,2*}, Jiewei Jiang^{3*} and
Wei Chen^{1,2*}

¹National Clinical Research Center for Ocular Diseases, Eye Hospital, Wenzhou Medical University, Wenzhou, China, ²Ningbo Eye Hospital, Wenzhou Medical University, Ningbo, China, ³School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China, ⁴Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China, ⁵Department of Ophthalmology, Wenzhou Hospital of Integrated Traditional Chinese and Western Medicine, Wenzhou, China

Purpose: To develop a visual function-based deep learning system (DLS) using fundus images to screen for visually impaired cataracts.

Materials and methods: A total of 8,395 fundus images (5,245 subjects) with corresponding visual function parameters collected from three clinical centers were used to develop and evaluate a DLS for classifying non-cataracts, mild cataracts, and visually impaired cataracts. Three deep learning algorithms (DenseNet121, Inception V3, and ResNet50) were leveraged to train models to obtain the best one for the system. The performance of the system was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity.

Results: The AUC of the best algorithm (DenseNet121) on the internal test dataset and the two external test datasets were 0.998 (95% CI, 0.996–0.999) to 0.999 (95% CI, 0.998–1.000), 0.938 (95% CI, 0.924–0.951) to 0.966 (95% CI, 0.946–0.983) and 0.937 (95% CI, 0.918–0.953) to 0.977 (95% CI, 0.962–0.989), respectively. In the comparison between the system and cataract specialists, better performance was observed in the system for detecting visually impaired cataracts ($p < 0.05$).

Conclusion: Our study shows the potential of a function-focused screening tool to identify visually impaired cataracts from fundus images, enabling timely patient referral to tertiary eye hospitals.

KEYWORDS

artificial intelligence, deep learning, visual impairment, cataracts, fundus images

Abbreviations: DLS; Deep learning system, ROC; Receiver operating characteristic curve, WHO; World Health Organization, LOCS; Lens Opacities Classification System, ZEHWZ; Zhejiang Eye Hospital at Wenzhou, ZEHHZ; Zhejiang Eye Hospital at Hangzhou, NEH; Ningbo Eye Hospital, BCDVA; Best corrected decimal visual acuity, CNN; Convolutional neural network, ADAM; Adaptive Estimation of Moments, t-SNE; t-distributed stochastic neighbour embedding; GradCAM, Gradient-weighted Class Activation Mapping; CIs, Confidence intervals; AUC, Area under the curve.

1 Introduction

Worldwide, the incidence of visual impairment is increasing (GBD, 2019 Blindness and Vision Impairment Collaborators, 2021), which is an important public health problem, with cataracts being the leading cause of visual impairment (Flaxman et al., 2017). According to recent research, among the 2.2 billion people who suffer from visual impairment worldwide, 134 million are blind, and 571 million have moderate-to-severe visual impairment in 2020 due to cataracts (Bourne et al., 2017; Flaxman et al., 2017). In low- and middle-income countries, especially in Southeast Asia and Africa, cataracts lead to higher rates of visual impairment than in high-income countries due to limited healthcare and financial resources (Lam et al., 2015). The World Health Organization (WHO) has adopted a 30 percent increase in effective coverage of cataract surgery as a new global target for eye care by 2030 (WHO, 2021). Therefore, there is an urgent need to facilitate and expedite cataract screening capabilities, especially for underserved populations.

Traditional cataract screening requires a professional ophthalmologist to assess the lens through a slit-lamp microscope (Gali et al., 2019) and grading methods based on the lens opacity classification system LOCS II (Chylack et al., 1989) or LOCS III (Chylack et al., 1993) (Lens Opacities Classification System, LOCS) and Wisconsin cataract grading system (Wong et al., 2013), which limits the efficiency of large-scale cataract screening. A simple and effective model for screening and referral remains a key challenge for the sustainable implementation of cataract screening programs. To enhance community screening for retinal disease in some countries (Lian et al., 2016; Verbraak et al., 2019), they have implemented telemedicine or artificial intelligence analysis of fundus images acquired by non-specialists. Grading the assessment of cataracts by fundus images may also be an effective solution. Abdul-Rahman used Fourier analysis to quantify optical degradation in fundus images, which was shown to be correlated well with the LOCS III (Abdul-Rahman et al., 2008).

Several studies have developed deep learning systems (DLSs) to grade the severity of cataracts based on the blurriness of fundus images. According to the visibility of the optic disk or retinal vessels of the fundus images, they classified cataracts into 3, four or 5 grades (Xiong et al., 2017; Zhang et al., 2019; Xu et al., 2020; Yue Zhou and Li, 2020). Considering that visual acuity is one of the most common indicators for evaluating the impact of cataracts on patients, it would be more meaningful to establish a visual function-based cataract grading system (WHO, 2020). This functional cataract screening program is more targeted for cataract patients, which can reduce the excessive referral of people with mild visual impairment and reduce the pressure on tertiary eye hospitals.

In this study, we developed a visual function-based DLS for populations based on fundus images, especially for the screening of visually impaired cataracts. In addition, we used images taken by different types of fundus cameras from three institutions to evaluate the effectiveness and generalizability of the system.

2 Materials and Methods

2.1 Image datasets

In this retrospectively study, a total of 6,997 fundus images (4,346 subjects) collected from Zhejiang Eye Hospital at Wenzhou (ZEHWZ) between September 2020 and March 2021 were used to develop the DLS. The ZEHWZ dataset included cataract patients whose best corrected decimal visual acuity (BCDVA) was good (>0.6) within 1 month after cataract surgery and non-cataract patients without refractive media opacities. The fundus images were captured without mydriasis before surgery. The exclusion criteria were traumatic cataracts, congenital cataracts and lens dislocation, corneal diseases, asteroid hyalosis, vitreous haemorrhage, and severe retinal and optic nerve diseases. Poor quality and unreadable images were also excluded: images out of focus; images underexposed; images overexposed; incomplete images with more than 1/3 peripheral halo.

Two additional datasets, including 1,398 fundus images obtained from two other institutions retrospectively, adopted the same inclusion criteria and exclusion criteria as ZEHWZ for external testing. One was derived from the inpatient department at Zhejiang Eye Hospital at Hangzhou (ZEHHZ), consisting of 1,097 images from 730 individuals; the other was derived from outpatient clinics and the inpatient department at Ningbo Eye Hospital (NEH), consisting of 301 images from 169 individuals.

This study adhered to the principles of the Declaration of Helsinki and was approved by the Ethics Committee of Zhejiang Eye Hospital at Wenzhou (Number, 2022-008-K-06-01). Due to the retrospective study design and the use of fully anonymized fundus images, the need for informed patient consent was waived by the review committee.

2.2 Criteria of cataract classification

The diagnosis of each fundus image was diagnosed by two cataract specialists based on the previous medical records and the results of the ophthalmology examination. If there was a difference between the two cataract specialists, there would be a third senior cataract specialists for diagnosis. All fundus images with a definitive diagnosis were screened for quality control. Poor quality and unrecognizable images were excluded.

All fundus images were classified into three categories: non-cataracts, mild cataracts, and visually impaired cataracts. Non-cataracts were defined as patients with transparent lenses and without refractive media opacities. Mild cataracts were defined as cataracts with mild vision impairment with $BCDVA \geq 0.3$, and visually impaired cataracts were defined as cataracts with moderate-to-severe vision impairment or blindness with $BCDVA < 0.3$. Typical examples of non-cataract and cataract fundus images are displayed in Figure 1.

2.3 Image preprocessing

During image preprocessing, each image was uniformly scaled down to 224×224 pixels, and the pixel values were normalized

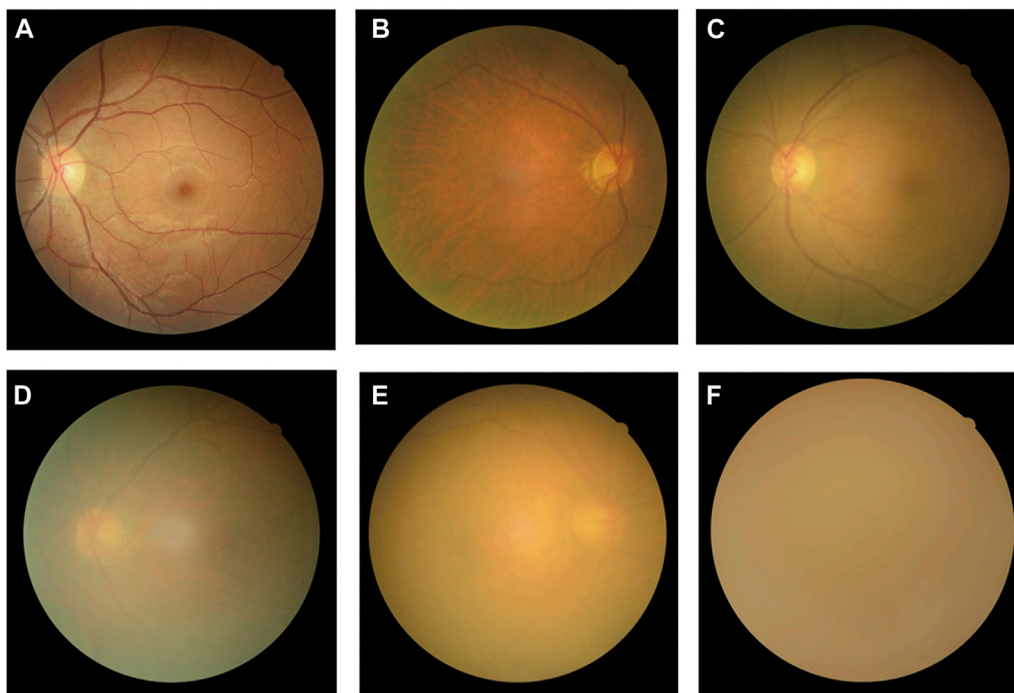


FIGURE 1
 Typical examples of fundus images of non-cataracts, mild cataracts, and visually impaired cataracts (A) Non-cataracts (B) The cataract with BCDVA = 0.8 (C) The cataract with BCDVA = 0.5 (D) The cataract with BCDVA = 0.3 (E) The cataract with BCDVA = 0.1 (F) The cataract with BCDVA = HM/BE.

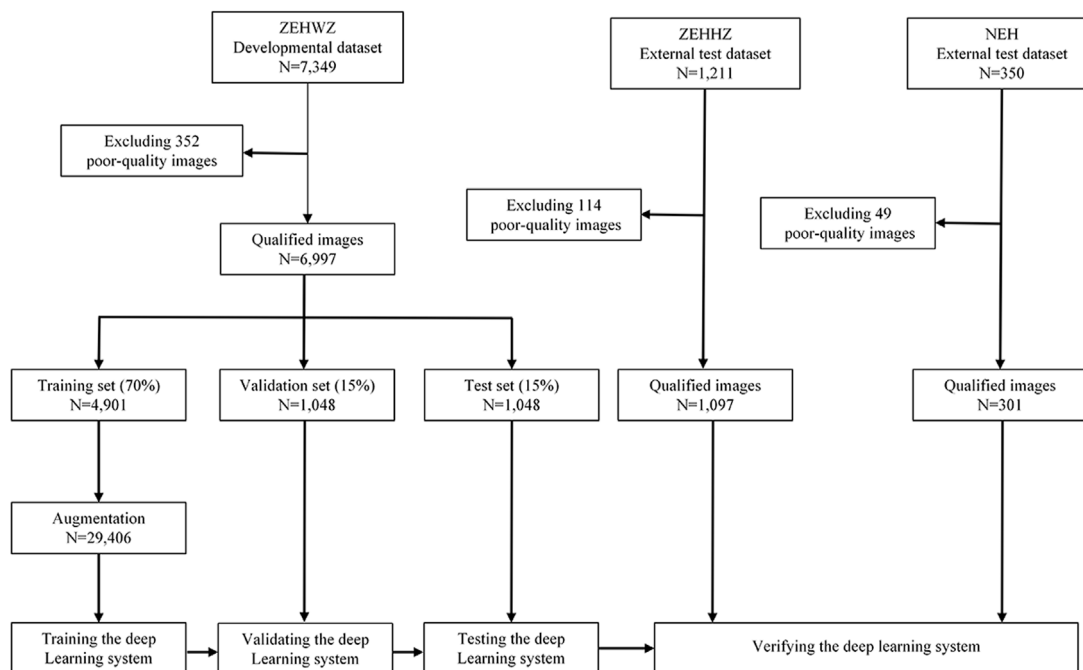


FIGURE 2
 Flow chart for the development and evaluation of the DLS. ZEHWZ = Zhejiang Eye Hospital at Wenzhou; ZEHHZ = Zhejiang Eye Hospital at Hangzhou; NEH = Ningbo Eye Hospital.

between 0 and 1. Then, data augmentation techniques were applied to increase the diversity of the dataset and thereby alleviate the overfitting problem during deep learning training. The new samples were generated by a simple transformation of the original image, simulating “real world” acquisition conditions. Random cropping, rotation of 90°, and horizontal and vertical flipping were applied to the images of the training dataset to increase the sample size to six times the original size (from 4,901 to 29,406).

2.4 Development and evaluation of the DLS

The fundus images drawn from the ZEHWZ dataset were randomly divided into training, validation, and internal test datasets at a ratio of 70%:15%:15%. The training and validation datasets were used to develop the system, and the test dataset was used to evaluate the performance of the system. Images from the same person were only assigned to a single dataset to prevent deep learning leaks and biased evaluations.

To find the best deep learning model for distinguishing non-cataracts, mild cataracts, and visually impaired cataracts, three convolutional neural network (CNN) architectures (DenseNet121, Inception-v3, and ResNet50) were compared. The parameters of the CNN were initialized with weights pretrained for ImageNet classification.

The deep learning models were trained using PyTorch (version 1.6.0) as the backend. Using the Adaptive Estimation of Moments (ADAM) optimizer, the initial learning rate was 0.001, β_1 was 0.9, β_2 was 0.999, and the weight decay was $1e-4$. Each model was trained for 80 epochs. During the training, the validation loss was evaluated on the validation dataset after each epoch and used as a reference for model selection. Each time the validation loss was reduced, the model state and corresponding weight matrix were saved. The model state with the lowest validation loss was saved as the final state of the model for the test dataset.

The diagnostic performance of the three-class classification model was then evaluated on two independent external test datasets. The development and evaluation process of the system is shown in [Figure 2](#). Using the t-distributed stochastic neighbour embedding (t-SNE) technique, the embedding features of each class learned by the model were displayed in a two-dimensional space.

2.5 Visualization heatmap

To understand which areas of fundus images were most likely to be used by deep learning models to generate decisions for this system, we use the Gradient-weighted Class Activation Mapping (GradCAM) technique to generate heatmaps. This technique uses the gradients of any target concept, flowing into the final convolutional layer to produce a localization map highlighting the important regions in the image for predicting the concept ([Ramprasaath et al., 2020](#)). Hotter colours represent the regions with more contribution to the predicted output, while cooler colours may indicate relatively less contribution to the predicted output. Using this method, heatmaps were generated to illustrate the basic principles of DLSs in differentiating between non-cataracts, mild cataracts, and visually impaired cataracts.

2.6 Characteristics of misclassification by the deep learning system

A senior cataract specialists who had not been involved in the initial diagnosis reviewed the characteristics of all images misclassified by the DenseNet121 algorithm and analysed the possible causes of misclassification in combination with the corresponding BCDVA.

2.7 DLS versus cataract specialists

To assess our DLS in the context of cataract detection, we recruited two cataract specialists with 3 and 10 years of clinical experience. The ZEHHZ dataset was employed to compare the performance of the best system (DenseNet121) to that of the cataract specialists with the reference standard. The system and specialists independently classified each image into one of the following three categories: non-cataracts, mild cataracts, and visually impaired cataracts. Notably, to reflect the level of experience of the cataract specialists in normal clinical practice, they were not told that they were competing with an AI-based system to avoid competition bias.

2.8 Statistical analysis

The performance of the deep learning system for the classification of non-cataracts, mild cataracts, and visually impaired cataracts was evaluated by employing the one-versus-rest tactic and calculating the AUC, sensitivity, specificity, and accuracy. Statistical analysis was performed using Python 3.7.8 (Wilmington, Delaware, United States of America). The 95% confidence intervals (CIs) for sensitivity, specificity, and accuracy were calculated by the Wilson scoring method using the Stats model package (version 0.11.1), and those for the area under the receiver operating characteristic (ROC) curve (AUC) were calculated using an empirical bootstrap procedure with 1,000 repetitions. We plotted the receiver operating characteristic (ROC) curve to demonstrate the capability of the system by plotting the ratio of true positive cases (sensitivity) to false positive cases (1-specificity) using the Scikit-learn (version 0.23.2) and Matplotlib (version 3.3.1) packages; a larger AUC indicated better performance. Unweighted Cohen's kappa coefficients were calculated to compare the results of the system to a reference standard. Differences in sensitivity, specificity, and accuracy between systems and the cataract specialists were analysed using the McNemar test. All statistical tests were two-sided with a significance level of 0.05.

3 Results

3.1 Characteristics of the datasets

After removing 515 poor-quality images, a total of 8,395 qualified images (3,569 images of non-cataracts, 3,245 images of mild cataracts, and 1,581 images of visually impaired cataracts) from 5,245 individuals were used to develop

TABLE 1 Summary of datasets.

Item	ZEHWZ dataset			ZEHHZ dataset	NEH dataset
Total no. of images	7,349			1,211	350
Total no. of qualified images	6,997			1,097	301
No. of subjects	4,346			730	169
Age, mean/range (years)	46.54/5–92			50.70/3–92	48.04/4–87
No. (%) of women	2,333/53.68			425/58.22	99/58.58
Camera model	Canon CR-2 PLUS AF (Japan)			Canon CR-2 (Japan)	RetiCam 3,100 (China)
	Training Set (70%) 4,901	Validation Set (15%) 1,048	Test Set (15%) 1,048		
Non-cataracts No. (%)	2,141 (43.68)	458 (43.70)	458 (43.70)	405 (36.92)	107 (35.55)
Mild cataracts No. (%)	1808 (36.89)	387 (36.93)	387 (36.93)	560 (51.05)	103 (34.22)
Visually impaired cataracts No. (%)	952 (19.42)	203 (19.37)	203 (19.37)	132 (12.03)	91 (30.23)

ZEHWZ = zhejiang eye hospital at wenzhou; ZEHHZ = zhejiang eye hospital at hangzhou; NEH, ningbo eye hospital.

and externally evaluate the DLS. Further information on the datasets from ZEHWZ, ZEHHZ, and NEH is summarized in Table 1.

3.2 Performance of different deep learning algorithms on the internal test dataset

This study used three classical deep learning algorithms, DenseNet121, ResNet50, and Inception-v3, to train the models. The t-SNE technique showed that the features of each category learned by the DenseNet121 algorithm were more separable than those learned by ResNet50 and Inception-v3 (Figure 3A). The performance of the three algorithms on the internal test dataset is shown in Figures 3B,C, which indicates that the best algorithm was DenseNet121. More information, including the accuracy, sensitivity, and specificity of the algorithms, is presented in Table 2.

The best algorithm achieved an AUC of 0.999 (95% confidence interval [CI], 0.998–1.000), a sensitivity of 98.3% (95% CI, 97.1–99.5), and a specificity of 98.8% (95% CI, 97.9–99.7) in detecting non-cataracts. The best algorithm discriminated mild cataracts from non-cataracts and visually impaired cataracts with an AUC of 0.958 (95% CI, 0.946–0.968), a sensitivity of 83.2% (95% CI, 79.5–86.9), and a specificity of 94.1% (95% CI, 92.3–95.9). The best algorithm discriminated visually impaired cataracts from non-cataracts and mild cataracts with an AUC of 0.956 (95% CI, 0.944–0.968), a sensitivity of 84.7% (95% CI, 79.8–89.7), and a specificity of 93.1% (95% CI, 91.4–94.8). Based on the reference standard of the internal test dataset, the unweighted Cohen's kappa coefficient of the best algorithm, DenseNet121, was 0.845 (0.817–0.873).

3.3 Performance of the different deep learning algorithms on the external test datasets

The performance of the DenseNet121, ResNet50, and Inception-v3 algorithms for cataract validation on the external test dataset is

shown in Figure 4, confirming that DenseNet121 achieved the best performance. The t-SNE technique also indicated that the features of each category learned by the DenseNet121 algorithm were more separable than those learned by Inception-v3 and ResNet50 (Figure 4A–D).

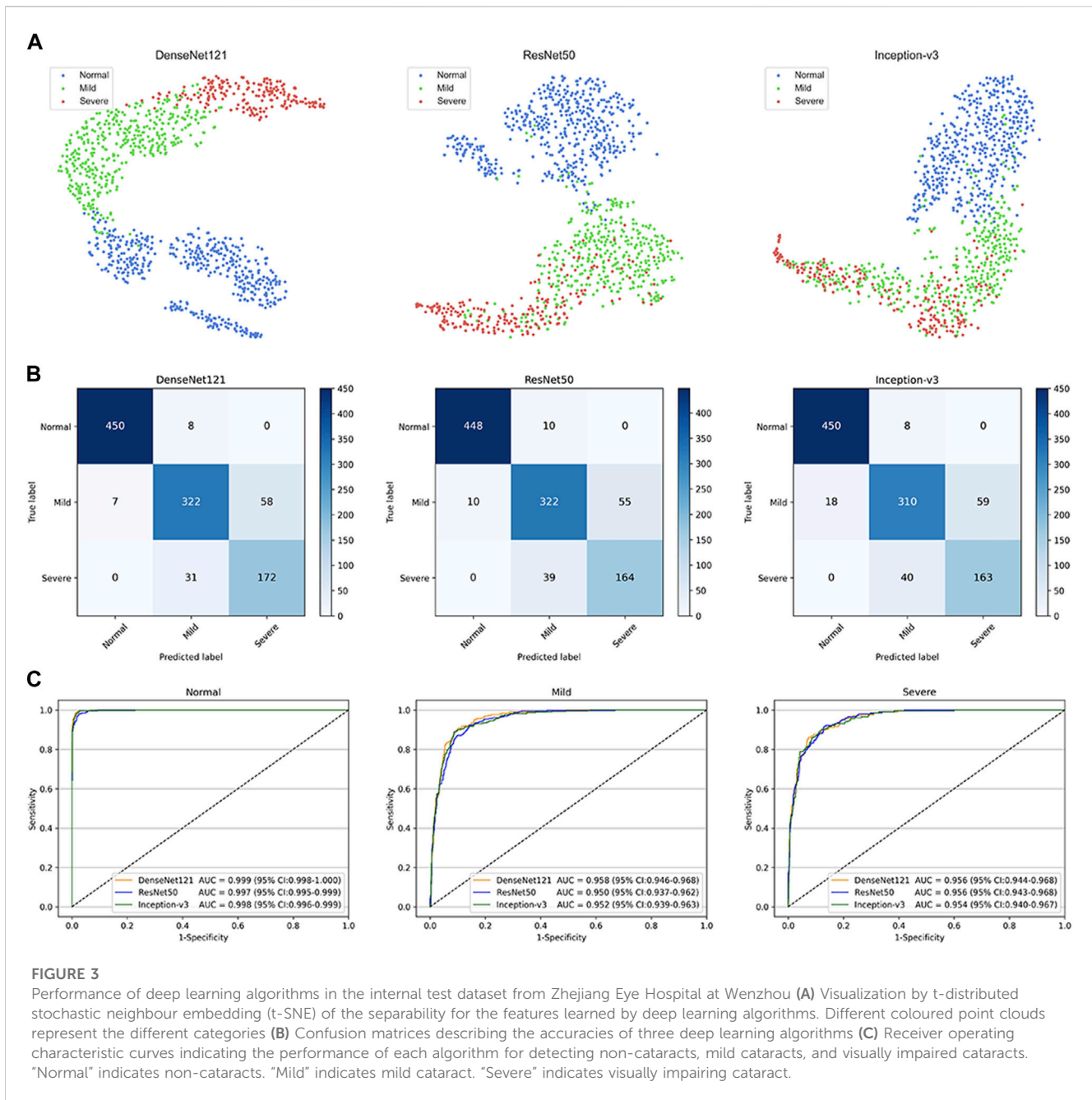
For the ZEHHZ dataset, the system based on DenseNet121 achieved AUCs of 0.998 (95% CI, 0.996–0.999), 0.938 (95% CI, 0.924–0.951), and 0.937 (95% CI, 0.918–0.953) in the classification of non-cataracts, mild cataracts, and visually impaired cataracts, respectively. In the NEH dataset, the system based on DenseNet121 achieved AUCs of 0.998 (95% CI, 0.995–1.000), 0.966 (95% CI, 0.946–0.983), and 0.977 (95% CI, 0.962–0.989) in the classification of non-cataracts, mild cataracts, and visually impaired cataracts, respectively.

The details on the classification performance of the three algorithms with the external datasets are shown in Table 2. In the ZEHHZ dataset, the accuracies of the best algorithm (DenseNet121) in the detection of non-cataracts, mild cataracts, and visually impaired cataracts were 97.3% (95% CI, 96.3–98.2), 85.5% (95% CI, 83.4–87.6), and 88.2% (95% CI, 86.3–90.1), respectively. In the NEH dataset, the accuracies of the best algorithm in the detection of non-cataracts, mild cataracts, and visually impaired cataracts were 98.7% (95% CI, 97.4–100.0), 89.7% (95% CI, 86.3–93.1), and 91.0% (95% CI, 87.8–94.3), respectively.

Based on the reference standards of the ZEHHZ and NEH datasets, the unweighted Cohen's kappa coefficients of the best algorithm, DenseNet121, were 0.762 (0.728–0.796) and 0.845 (0.793–0.897), respectively.

3.4 Heatmaps

We use heatmaps to provide insights into regions of the fundus images that might influence the algorithm's prediction. Based on the heatmaps shown in Figure 5, we observed that the regions highlighted by the algorithm matched well with the clear features on the fundus image. For the fundus images of the non-cataracts, the region highlighted by the heatmaps was relatively consistent: large



range, circular, and centred. For the fundus images of mild cataracts, the regions highlighted by the heatmaps are smaller, eccentric, oval, and around the optic disk, For the fundus images of visually impaired cataracts, the regions highlighted by the heatmaps are irregular. Figure 5 shows typical heatmaps of non-cataracts, mild cataracts, and visually impaired cataracts, respectively.

3.5 Classification errors

In the internal and external test datasets, a total of 293 images (11.98% of the total 2,446) were inconsistent with the diagnostic reference standard by the DenseNet121 algorithm. In the non-cataracts group (970 images), 38 images (3.92%) were

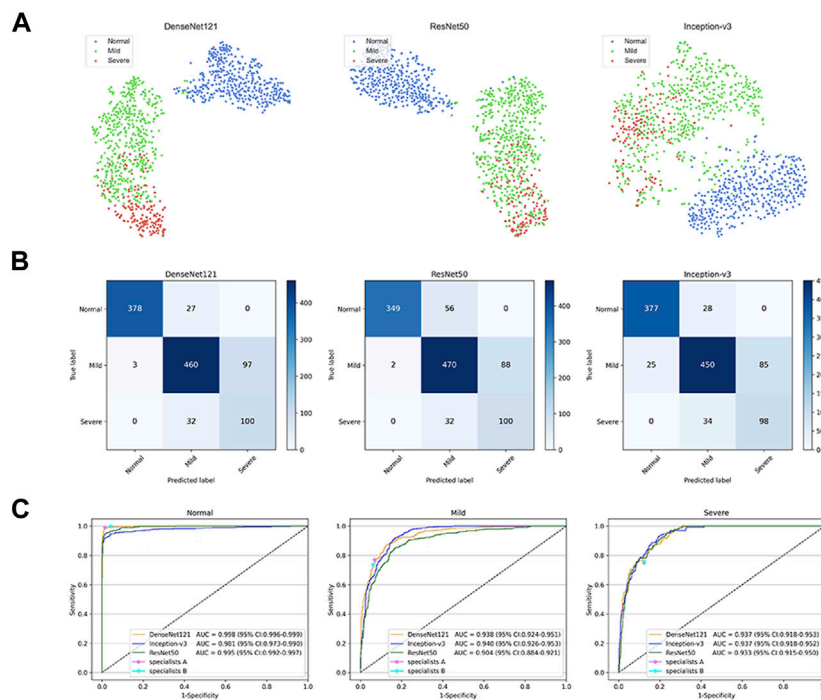
misclassified as mild cataracts by the system, 89.47% (34 images) of which were misclassified due to dark shooting, the region highlighted by the heatmaps was eccentric and oval, as the mild cataracts, for the images were slightly darker, slightly defocused or surrounded by the halo. In the mild cataracts group (1,050 images), 11 images (1.05%) were misclassified as non-cataracts by the system due to clarity of the fundus images, most of the patients are early cortical or nuclear cataracts, the highlighted region of the heatmaps show large range, circular, and centred, as the non-cataracts. 167 (15.90%) images were misclassified as visually impaired cataracts by the system, of which 65.27% images had relatively poor BCDVA (BCDVA < 0.5) with blurred fundus images and 10.78% had good BCDVA (BCDVA between 0.8–1.0) with advanced cortical opacity, whose fundus images were blurred, the highlighted region of the

TABLE 2 Performance of three deep learning algorithms in the internal and external test datasets.

One-vs.-rest classification	ZEHWZ internal test dataset			ZEHHZ external test dataset			NEH external test dataset		
	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Normal vs mild + severe									
DenseNet121	98.3% (97.1–99.5)	98.8% (97.9–99.7)	98.6% (97.8–99.3)	93.3% (90.9–95.8)	99.6% (99.1–100.0)	97.3% (96.3–98.2)	97.2% (94.1–100.0)	99.5% (98.5–100.0)	98.7% (97.4–100.0)
ResNet50	97.8% (96.5–99.2)	98.3% (97.3–99.3)	98.1% (97.3–98.9)	86.2% (82.8–89.5)	99.7% (99.3–100.0)	94.7% (93.4–96.0)	96.3% (92.7–99.9)	99.0% (97.5–100.0)	98.0% (96.4–99.6)
Inception-v3	98.3% (97.1–99.5)	96.9% (95.6–98.3)	97.5% (96.6–98.5)	93.1% (90.6–95.6)	96.4% (95.0–97.8)	95.2% (93.9–96.4)	94.4% (90.0–98.8)	92.3% (88.5–96.0)	93.0% (90.1–95.9)
Mild vs normal + severe									
DenseNet121	83.2% (79.5–86.9)	94.1% (92.3–95.9)	90.1% (88.3–91.9)	82.1% (79.0–85.3)	89.0% (86.4–91.7)	85.5% (83.4–87.6)	87.4% (81.0–93.8)	90.9% (86.9–94.9)	89.7% (86.3–93.1)
ResNet50	83.2% (79.5–86.9)	92.6% (90.6–94.6)	89.1% (87.2–91.0)	83.9% (80.9–87.0)	83.6% (80.5–86.7)	83.8% (81.6–86.0)	88.3% (82.2–94.5)	88.9% (84.5–93.3)	88.7% (85.1–92.3)
Inception-v3	80.1% (76.1–84.1)	92.7% (90.8–94.7)	88.1% (86.1–90.0)	80.4% (77.1–83.6)	88.5% (85.8–91.2)	84.3% (82.2–86.5)	72.8% (64.2–81.4)	88.4% (83.9–92.8)	83.1% (73.3–89.3)
Severe vs normal + mild									
DenseNet121	84.7% (79.8–89.7)	93.1% (91.4–94.8)	91.5% (89.8–93.2)	75.8% (68.4–83.1)	89.9% (88.1–91.8)	88.2% (86.3–90.1)	83.5% (75.9–91.1)	94.3% (91.1–97.4)	91.0% (87.8–94.3)
ResNet50	80.3% (74.8–85.8)	93.0% (91.3–94.7)	90.6% (88.8–92.3)	74.2% (66.8–81.7)	91.2% (89.4–93.0)	89.2% (87.3–91.0)	81.3% (73.3–89.3)	93.8% (90.6–97.1)	90.0% (86.6–93.4)
Inception-v3	80.8% (75.4–86.2)	93.5% (91.8–95.2)	91.0% (89.3–92.8)	75.8% (68.4–83.1)	90.9% (89.1–92.7)	89.1% (87.2–90.9)	80.2% (72.0–88.4)	95.2% (92.4–98.1)	90.7% (87.4–94.0)

ZEHWZ = zhejiang eye hospital at wenzhou; ZEHHZ = zhejiang eye hospital at hangzhou; NEH = ningbo eye hospital. "Normal" indicates non-cataracts. "Mild" indicates mild cataracts. "Severe" indicates visually impaired cataracts.

ZEHHZ external test dataset



NEH external test dataset

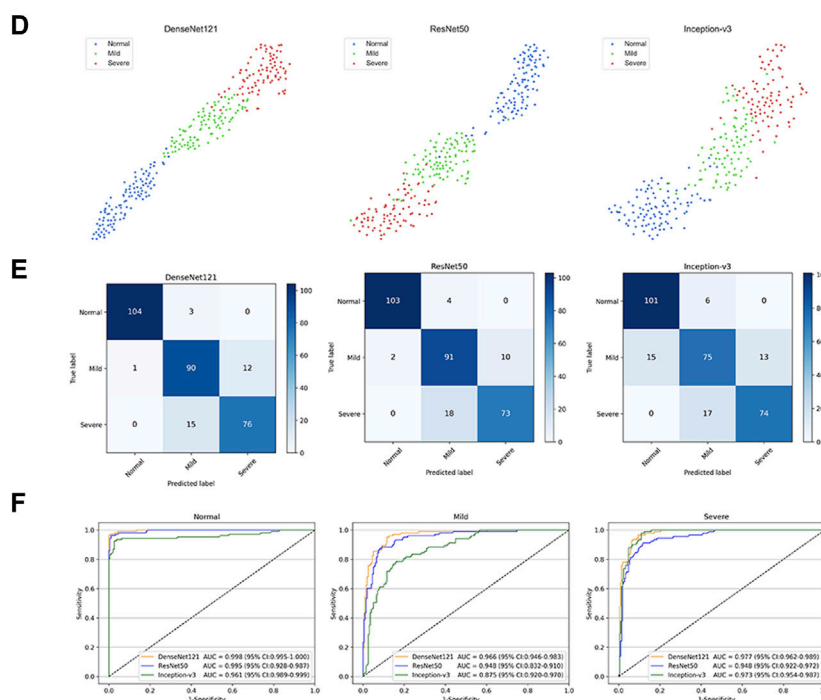


FIGURE 4

Confusion matrices and receiver operating characteristic (ROC) curves for three deep learning algorithms performance in two external test datasets. The t-distributed stochastic neighbour embedding (t-SNE) (A–D) presenting the separability for the features learned by deep learning algorithms in ZEHHZ and NEH external test datasets. Confusion matrices (B–E) describing the accuracies of two deep learning algorithms in the ZEHHZ and NEH external test datasets. ROC curves (C–F) indicating the performance of each algorithm for discriminating among non-cataracts, mild cataracts, and visually impaired cataracts in the ZEHHZ and NEH external test datasets. The performance of two cataract specialists were also indicated (C). ZEHHZ, Zhejiang Eye Hospital at Hangzhou. NEH, Ningbo Eye Hospital. “Normal” indicates non-cataracts. “Mild” indicates mild cataract. “Severe” indicates visually impaired cataract.

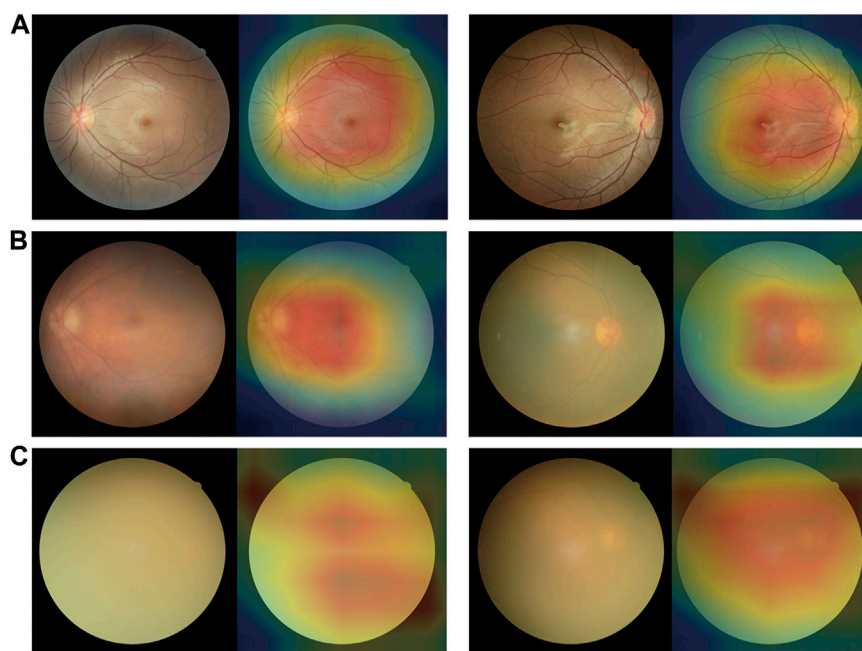
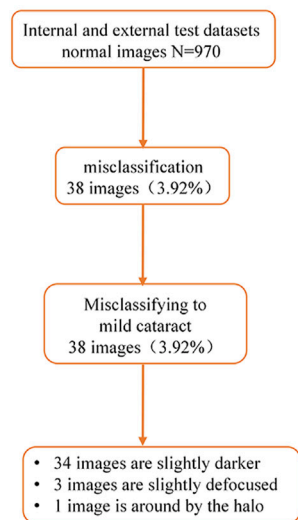
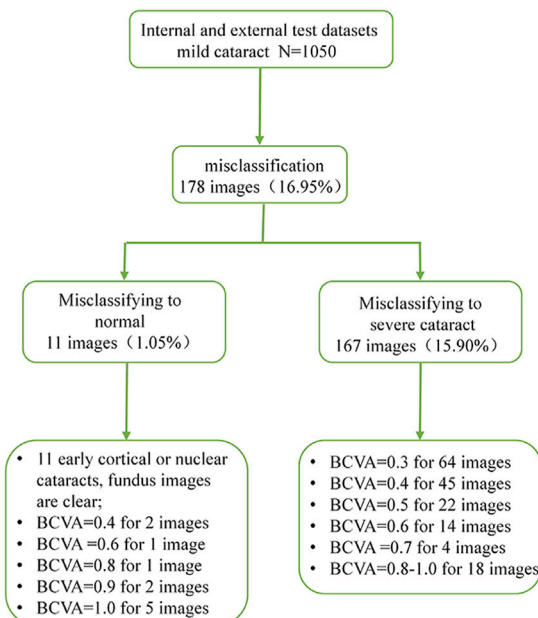


FIGURE 5 Saliency maps highlighting regions that the algorithm focuses on when making classification **(A)** Non-cataracts **(B)** mild cataracts **(C)** visually impaired cataracts. Each category is shown in a pair of an original image (left) and a corresponding heatmap (right). In these heatmaps, hotter areas (i.e., reds and oranges) are indicative of regions with increased contributions towards the predicted output, and colder regions (blues and greens) might be indicative of relatively less contribution. For each subgroup, each set of two images (from two different eyes) consistently shows the same region or feature highlighted by the algorithm.

A The misclassification of "normal"



B The misclassification of "mild cataract"



C The misclassification of "severe cataract"

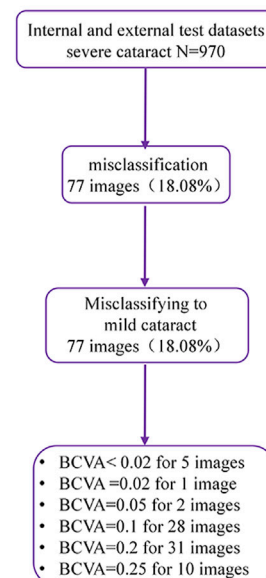


FIGURE 6 Details of deep learning system error classification in internal and external test datasets. **(A)** The misclassification of the non-cataracts group; **(B)** The misclassification of the mild cataracts group; **(C)** The misclassification of the visually impaired cataracts group.

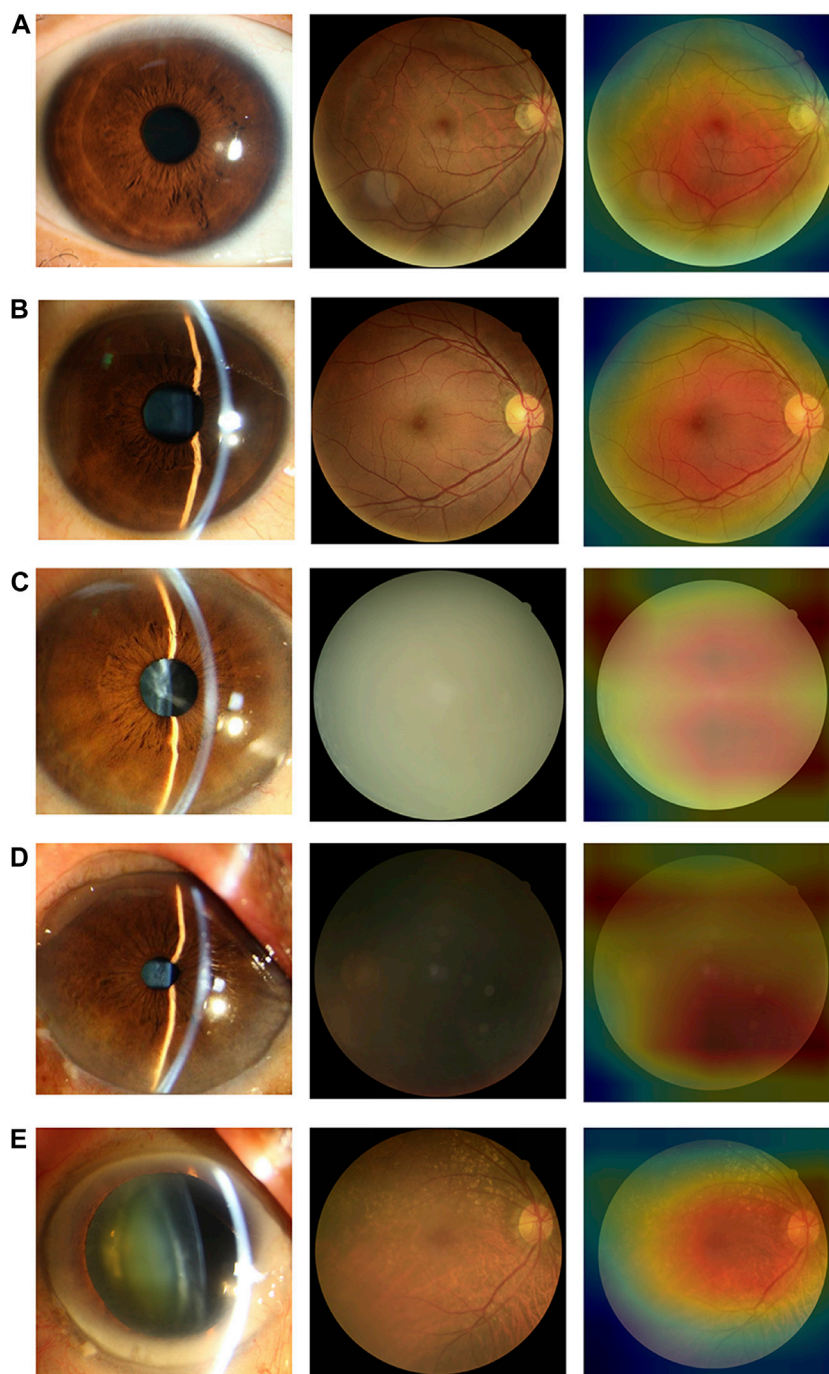


FIGURE 7

Typical examples of misclassified images by the DLS **(A)** Images of “non-cataract” incorrectly classified as “mild cataract”. The fundus image was around by the halo **(B)** Images of “mild cataract” incorrectly classified as “non-cataract”. The patient had cataracts in the early stage, BCDVA = 1.0 **(C)** Images of “mild cataract” incorrectly classified as “visually impaired cataract”. The patient had advanced cortical opacity, BCDVA = 0.6 **(D)** Images of “mild cataract” incorrectly classified as “visually impaired cataract”. Patients with small pupils reduced the amount of light entering their eyes (BCDVA = 0.4) **(E)** Images of “visually impaired cataract” incorrectly classified as “mild cataract”. The patient had a small-scale posterior subcapsular area, BCDVA = 0.16.

heatmaps was irregular, as the visually impaired cataracts. In the visually impaired cataracts group (426 images), 77 images (18.08%) were systematically misclassified as mild cataracts, the heatmaps show the characteristic of the mild cataracts: smaller, eccentric, oval, and around the optic disk, because among these classification errors,

most cataracts' BCDVAs were not too bad (89.61% of the Images had BCDVA ≥ 0.1). The misclassification BCDVA situation of the DLS is shown in [Figure 6](#). [Figure 7](#) shows typical example of misclassified images of “non-cataract” incorrectly classified as “mild cataract”, misclassified images of “mild cataract”

TABLE 3 Performance comparison of DenseNet121 with cataract specialists in the ZEHHZ dataset.

	DenseNet121	Specialists A	Specialists B	P1	P2
Normal vs mild + severe					
Sensitivity (95% CI)	93.3% (90.9–95.8)	99.0% (98.0–100.0)	99.8% (99.3–100.0)	0.000	0.000
Specificity (95% CI)	99.6% (99.1–100.0)	98.6% (97.7–99.4)	95.8% (94.3–97.3)	0.065	0.000
Accuracy (95% CI)	97.3% (96.3–98.2)	98.7% (98.1–99.4)	97.3% (96.3–98.2)	0.014	1.000
Mild vs normal + severe					
Sensitivity (95% CI)	82.1% (79.0–85.3)	77.0% (73.5–80.5)	73.6% (69.9–77.2)	0.001	0.000
Specificity (95% CI)	89.0% (86.4–91.7)	93.1% (91.0–95.3)	93.9% (91.8–95.9)	0.002	0.000
Accuracy (95% CI)	85.5% (83.4–87.6)	84.9% (82.7–87.0)	83.5% (81.3–85.7)	0.576	0.074
Severe vs normal + mild					
Sensitivity (95% CI)	75.8% (68.4–83.1)	75.0% (67.6–82.4)	75.8% (68.4–83.1)	1.000	1.000
Specificity (95% CI)	89.9% (88.1–91.8)	87.7% (85.6–89.7)	87.7% (85.6–89.7)	0.005	0.006
Accuracy (95% CI)	88.2% (86.3–90.1)	86.1% (84.1–88.2)	86.2% (84.2–88.3)	0.012	0.019

ZEHHZ = Zhejiang Eye Hospital at Hangzhou. P1 refers to the *p*-value that was calculated between the deep learning system and cataract specialist A using the two-sided McNemar test. P2 refers to the *p*-value that was calculated between the deep learning system and cataract specialist B using the two-sided McNemar test. Cataract specialist A has 3 years of clinical experience. Cataract specialist B has 10 years of clinical experience. “Normal” indicates non-cataracts. “Mild” indicates mild cataract. “Severe” indicates visually impairing cataract.

incorrectly classified as “non-cataract”, images of “mild cataract” incorrectly classified as “visually impaired cataract”, and images of “visually impaired cataract” incorrectly classified as “mild cataract”, respectively.

3.6 Comparison of the deep learning system and cataract specialists

In the ZEHHZ dataset, for the classification of non-cataracts, mild cataracts, and visually impaired cataracts, the cataract specialist with 3 years of experience achieved accuracies of 98.7% (98.1–99.4), 84.9% (82.7–87.0), and 86.1% (84.1–88.2), respectively, the senior cataract specialist with 10 years of experience achieved accuracies of 97.3% (96.3–98.2), 83.5% (81.3–85.7) and 86.2% (84.2–88.3), respectively, and the DLS achieved accuracies of 97.3% (96.3–98.2), 85.5% (83.4–87.6) and 88.2% (86.3–90.1), respectively. Our system had comparable performance to that of cataract specialists in classifying non-cataracts and mild cataracts and had better performance in classifying visually impaired cataracts ($p < 0.05$) (Table 3 and Figure 4C).

4 Discussion

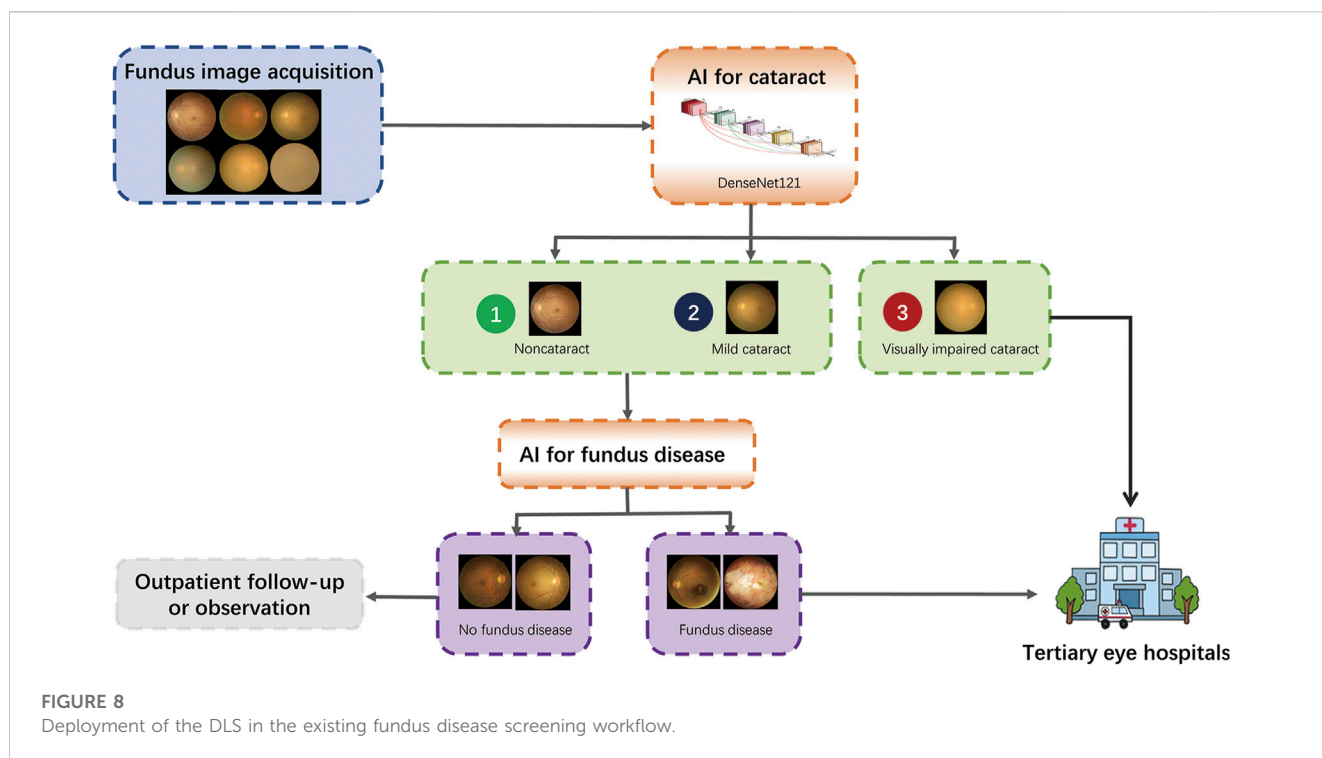
We developed a single-modality DLS using only fundus images to detect both mild cataracts and visually impaired cataracts in the general population. Our main finding was that the system based on a convolutional neural network could discriminate among non-cataracts, mild cataracts, and visually impaired cataracts, and the DenseNet121 algorithm had the best performance. In the internal and two external test datasets, the AUCs of the system based on the best algorithm were 0.998–0.999, 0.938–0.966, and 0.937–0.977, respectively, which demonstrated the broad generalizability of our system. In addition, the unweighted Cohen’s kappa coefficients were 0.762–0.845, which showed good consistency between the outcomes of the DLS and the reference standard,

further substantiating the effectiveness of our system. Moreover, our system has better performance in classifying visually impaired cataracts than cataract specialists.

The visual function-centric DLS in this study can serve as a simple, automated, and comprehensive cataract screening deployment tool. This system only needs to input fundus images and does not require other time-consuming and labour-intensive professional ophthalmic operations to obtain the severity of the patients’ cataract and the range of the best corrected visual acuity. Its simplicity can be used as an effective tool for community screening options, especially in resource-poor regions. It can not only screen for cataracts and but also can tell patients about their eye health. Moreover, visually impaired cataracts can be screened out and referred to tertiary eye hospitals.

With the increase in fundus disease-based primary care programs and community screening programs (Lin et al., 2021; Ruamviboonsuk et al., 2022), fundus photography is a routine examination procedure, and the cataract algorithm of this study can be used as an add-on algorithm to these existing devices with minimal additional cost to achieve more disease screening functions. In addition, the blurring of some fundus images caused by severe cataracts is a common cause of ungradable fundus disease (Scanlon et al., 2005). Our algorithm can screen out the fundus images of non-cataracts and mild cataracts because the fundus images of these two groups have relatively high definition, which can improve the accuracy of intelligent screening of fundus diseases and reduce the burden of unnecessary manual classification, enabling more effective referrals and improving the capacity of the existing screening programs for eye diseases. The visually impaired cataracts selected by the algorithm can be referred to a tertiary eye hospital for treatment. The workflow is shown in Figure 8.

Most of the previous studies on deep learning algorithms for cataracts based on fundus images focused on the artificial classification of the blurriness of the fundus images (Xiong et al., 2017; Zhang et al., 2019; Xu et al., 2020; Yue Zhou and Li, 2020). The annotations are subjective, and there is no accurate corresponding clinical guiding significance. In these studies, the application of these



algorithms did not meet the actual situation and needs of the communities, and most of the previous studies did not consider the state of visual function. Recently, [Tham et al. \(2022\)](#) developed an algorithm for the automatic detection of visually significant cataracts with an AUC of 0.916–0.966. However, their algorithm can only distinguish visually significant cataracts from mild cataracts in cataract patients, but our algorithm can further classify non-cataracts from cataracts, which is of great significance for cataract screening and eye health guidance in communities. At the same time, our algorithm can also distinguish mild cataracts from non-cataracts. Although the patients only need regular follow-up and observation, we can give them some suggestions for controlling and delaying the progression of cataracts, for numerous studies had found that the risk factors for cataract formation had been associated with lifestyle and systemic diseases, include smoking, ultra-violet light exposure, alcohol intake, nutritional status, diabetes mellitus, hypertension, obesity, chronic kidney disease and autoimmune disease ([Ang and Afshari, 2021](#)). Therefore, we can advise the patients to choose a healthy lifestyle and control systemic diseases, such as controlling blood sugar well. In addition, in our research, we compared three different CNN algorithms: DenseNet121, ResNet50, and Inception-v3. Among them, Densenet121 is the most accurate algorithm. It has a variety of advantages used in their study when compared to two other algorithms: alleviating the vanishing-gradient problem, strengthening feature propagation, encouraging feature reuse, and substantially improving parameter efficiency ([Huang et al., 2019](#)).

Reducing false negative misclassification of visually impaired cataracts is critical to avoid missing cataract patients who should be referred to tertiary eye centres for surgical intervention. A total of 18.08% (77/970) of visually impaired cataracts were misclassified as mild cataracts. Analysis of the misclassified fundus images found

that 89.61% (69/77) of them had moderate visual impairment ($0.1 \leq \text{BCDVA} < 0.3$). The optometry to get BCDVA is subjective and requires the patient's cooperation. Some cataract patients with relatively poor visual acuity might give up their efforts to see some small optotypes. Therefore, the actual visual acuity of the patients may be slightly better than the checked visual acuity. Additionally, this misclassification may be caused by a small-scale posterior subcapsular cataract. This type of cataract has a greater impact on visual acuity, while its small-scale turbidity has less impact on the quality of fundus images ([Stifter et al., 2005](#)). Reducing false positive cataract results for visually impaired cataracts is also an important consideration in community screening programs to avoid unnecessary referrals. In this study, 65.27% (109/178) of patients incorrectly diagnosed with cataracts had $\text{BCDVA} < 0.5$. In some countries, the population in need of cataract surgery is defined as having $\text{BCDVA} < 0.5$, with cataracts as the main cause of vision impairment or blindness ([WHO, 2021](#)). Referral of these patients would not waste medical resources. Some patients with advanced cortical opacity have poor contrast sensitivity, although their visual acuity is good ([Maraini et al., 1994](#)). Therefore, these false positives may still need to be referred to a tertiary eye centre and cannot be completely considered incorrect referrals.

This study has several limitations. First, we did not investigate the influence of corneal diseases and vitreous haemorrhage on fundus images. However, the incidence of spontaneous vitreous haemorrhage and corneal opacity in the general population is low, 0.007% ([Manuchehri and Kirkby, 2003](#)) and 3.7% ([Mukhija et al., 2020](#)), respectively. If the patient has corneal opacity or vitreous haemorrhage, he or she must go to the hospital for further examination, and the recommendation given by the system would still apply. Second, the optometry is affected by patient

compliance. Therefore, misclassification due to subjective measurement errors cannot be completely ruled out.

We developed and evaluated a novel single-modality, fundus image-based DLS for the detection of cataracts, especially visually impaired cataracts. The performance of the DLS is comparable to that of the experienced cataract specialist, indicating that this DLS can not only be used to screen cataract patients but also facilitate a timelier and more accurate referral of visually impaired cataract patients from communities to tertiary eye hospitals.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by This study adhered to the principles of the Declaration of Helsinki and was approved by the Ethics Committee of Zhejiang Eye Hospital at Wenzhou (Number, 2022-008-K-06-01). Due to the retrospective study design and the use of fully anonymized fundus images, the need for informed patient consent was waived by the review committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

Conception and design: HX, ZL, QZ, JJ, and WC. Funding obtainment: WC. Provision of study data: WC and HX. Collection

and assembly of data: CW, ZW, CL, QG, and MW. Data analysis and interpretation: HX, ZL, JJ, CW, WC, and YZ. Manuscript writing: All authors contributed to the article and approved the submitted version.

Funding

This study received funding from the National Key R&D Programme of China (grant no. 2019YFC0840708), the National Natural Science Foundation of China (grant no. 81970770), the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2019KY466), the National Natural Science Foundation of China (grant no. 62276210), the Natural Science Basic Research Program of Shaanxi (grant no. 2022JM-380) and the Wenzhou Science and Technology Foundation (grant no. Y20211005). The funding organizations played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abdul-Rahman, A. M., Molteno, T., and Molteno, A. C. (2008). Fourier analysis of digital retinal images in estimation of cataract severity. *Clin. Exp. Ophthalmol.* 36 (7), 637–645. doi:10.1111/j.1442-9071.2008.01819.x
- Ang, M. J., and Afshari, N. A. (2021). Cataract and systemic disease: A review. *Clin. Exp. Ophthalmol.* 49 (2), 118–127. doi:10.1111/ceo.13892
- Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., et al. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. *Lancet Glob. health* 5 (9), e888–e897. doi:10.1016/S2214-109X(17)30293-0
- Chylack, L. T., Jr., Leske, M. C., McCarthy, D., Khu, P., Kashiwagi, T., and Sperduto, R. (1989). Lens opacities classification system II (LOCS II). *Archives Ophthalmol.* 107 (7), 991–997. doi:10.1001/archophth.1989.01070020053028
- Chylack, L. T., Jr., Wolfe, J. K., Singer, D. M., Leske, M. C., Bullimore, M. A., Bailey, I. L., et al. (1993). The lens opacities classification system III. The longitudinal study of cataract study group. *Archives Ophthalmol.* 111 (6), 831–836. doi:10.1001/archophth.1993.01090060119035
- Flaxman, S. R., Bourne, R. R. A., Resnikoff, S., Ackland, P., Braithwaite, T., Cicinelli, M. V., et al. (2017). Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. *Lancet Glob. health* 5 (12), e1221–e1234. doi:10.1016/S2214-109X(17)30393-5
- Gali, H. E., Sella, R., and Afshari, N. A. (2019). Cataract grading systems: A review of past and present. *Curr. Opin. Ophthalmol.* 30 (1), 13–18. doi:10.1097/ICU.0000000000000542
- GBD 2019 Blindness and Vision Impairment Collaborators (2021). Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the global burden of disease study. *Lancet Glob. health* 9 (2), e130–e143. doi:10.1016/S2214-109X(20)30425-3
- Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., and Weinberger, K. (2019). Convolutional networks with dense connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 8704–8716. doi:10.1109/TPAMI.2019.2918284
- Lam, D., Rao, S. K., Ratra, V., Liu, Y., Mitchell, P., King, J., et al. (2015). *Cataract. Nat. Rev. Dis. Prim.* 1, 15014. doi:10.1038/nrdp.2015.14
- Lian, J. X., Gangwani, R. A., McGhee, S. M., Chan, C. K., Lam, C. L., Primary Health Care, G., et al. (2016). Systematic screening for diabetic retinopathy (DR) in Hong Kong: Prevalence of DR and visual impairment among diabetic population. *Br. J. Ophthalmol.* 100 (2), 151–155. doi:10.1136/bjophthalmol-2015-307382
- Lin, D., Xiong, J., Liu, C., Zhao, L., Li, Z., Yu, S., et al. (2021). Application of comprehensive artificial intelligence retinal Expert (CARE) system: A national real-world evidence study. *Lancet Digit. Health* 3 (8), e486–e495. doi:10.1016/S2589-7500(21)00086-8

- Manuchehri, K., and Kirkby, G. (2003). Vitreous haemorrhage in elderly patients: Management and prevention. *Drugs Aging* 20 (9), 655–661. doi:10.2165/00002512-200320090-00003
- Maraini, G., Rosmini, F., Graziosi, P., Tomba, M. C., Bonacini, M., Cotichini, R., et al. (1994). Influence of type and severity of pure forms of age-related cataract on visual acuity and contrast sensitivity. Italian American Cataract Study Group. *Invest. Ophthalmol. Vis. Sci.* 35 (1), 262–267.
- Mukhija, R., Gupta, N., Vashist, P., Tandon, R., and Gupta, S. K. (2020). Population-based assessment of visual impairment and pattern of corneal disease: Results from the CORE (corneal opacity rural epidemiological) study. *Br. J. Ophthalmol.* 104 (7), 994–998. doi:10.1136/bjophthalmol-2019-314720
- Ramprasaath, R., Selvaraju, M. C., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359. doi:10.1007/s11263-019-01228-7
- Ruamviboonsuk, P., Tiwari, R., Sayres, R., Nganthavee, V., Hemarat, K., Kongprayoon, A., et al. (2022). Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study. *Lancet Digit. Health* 4 (4), e235–e244. doi:10.1016/S2589-7500(22)00017-6
- Scanlon, P. H., Foy, C., Malhotra, R., and Aldington, S. J. (2005). The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. *Diabetes Care* 28 (10), 2448–2453. doi:10.2337/diacare.28.10.2448
- Stifter, E., Sacu, S., Benesch, T., and Weghaupt, H. (2005). Impairment of visual acuity and reading performance and the relationship with cataract type and density. *Invest. Ophthalmol. Vis. Sci.* 46 (6), 2071–2075. doi:10.1167/iovs.04-0890
- Tham, Y. C., Goh, J. H. L., Anees, A., Lei, X., Rim, T. H., Chee, M. L., et al. (2022). Detecting visually significant cataract using retinal photograph-based deep learning. *Nat. Aging* 2, 264–271. doi:10.1038/s43587-022-00171-6
- Verbraak, F. D., Abramoff, M. D., Bausch, G. C. F., Klaver, C., Nijpels, G., Schlingemann, R. O., et al. (2019). Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care* 42 (4), 651–656. doi:10.2337/dc18-0148
- WHO (2020). Discussion Paper Proposed global targets for 2030 on integrated people-centred eye care World Health Organization. Available at: https://cdn.who.int/media/docs/default-source/blindness-and-visual-impairment/discussion-paper-eye-care-indicators-and-proposed-targets-11-11-20.pdf?sfvrsn=ff4dad2_2.
- WHO (2021). Integrated people-centred eye care, including preventable vision impairment and blindness Global targets for 2030. World Health Organization Available at: https://apps.who.int/gb/ebwha/pdf_files/WHA74/A_9-en.pdf.
- Wong, W. L., Li, X., Li, J., Cheng, C. Y., Lamoureux, E. L., Wang, J. J., et al. (2013). Cataract conversion assessment using lens opacity classification system III and Wisconsin cataract grading system. *Invest. Ophthalmol. Vis. Sci.* 54 (1), 280–287. doi:10.1167/iovs.12-10657
- Xiong, L., Li, H., and Xu, L. (2017). An approach to evaluate blurriness in retinal images with vitreous opacity for cataract diagnosis. *J. Healthc. Eng.* 2017, 5645498. doi:10.1155/2017/5645498
- Xu, X., Zhang, L., Li, J., Guan, Y., and Zhang, L. (2020). A hybrid global-local representation CNN model for automatic cataract grading. *IEEE J. Biomed. Health Inf.* 24 (2), 556–567. doi:10.1109/JBHI.2019.2914690
- Yue Zhou, G. L., and Li, H. (2020). Automatic cataract classification using deep neural network with discrete state transition. *IEEE Trans. Med. Imaging* 39 (2), 436–446. doi:10.1109/TMI.2019.2928229
- Zhang, H., Niu, K., Xiong, Y., Yang, W., He, Z., and Song, H. (2019). Automatic cataract grading methods based on deep learning. *Comput. methods programs Biomed.* 182, 104978. doi:10.1016/j.cmpb.2019.07.006