Article

Detection of Referable Horizontal Strabismus in Children's Primary Gaze Photographs Using Deep Learning

Ce Zheng[1,2], Qian Yao[5], Jiewei Lu[3], Xiaolin Xie[4], Shibin Lin[4], Zilei Wang[2], Siyin Wang[2], Zhun Fan[3], and Tong Qiao[2]

[1] Department of Ophthalmology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[2] Department of Ophthalmology, Shanghai Children's Hospital, Shanghai Jiaotong University, Shanghai, China
[3] Department of Electronic Engineering, Shantou University, Shantou, Guangdong, China
[4] Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong, Shantou University Medical College, Shantou, Guangdong, China
[5] Department of Ophthalmology, Zhaoqing Gaoyao People's Hospital, Zhaoqing, Guangdong, China

Correspondence: Tong Qiao, Department of Ophthalmology, Shanghai Children's Hospital, Shanghai Jiao Tong University, Lu Ding Road # 355, Shanghai 200012, China.
e-mail: qiaojoel@163.com
Zhun Fan, Guangdong Provincial Key Laboratory of Digital Signal and Image Processing, College of Engineering, Shantou University, Shantou 515063, China.
zfan@stu.edu.cn

Purpose: This study implements and demonstrates a deep learning (DL) approach for screening referable horizontal strabismus based on primary gaze photographs using clinical assessments as a reference. The purpose of this study was to develop and evaluate deep learning algorithms that screen referable horizontal strabismus in children's primary gaze photographs.

Methods: DL algorithms were developed and trained using primary gaze photographs from two tertiary hospitals of children with primary horizontal strabismus who underwent surgery as well as orthotropic children who underwent routine refractive tests. A total of 7026 images (3829 non-strabismus from 3021 orthoptics [healthy] subjects and 3197 strabismus images from 2772 subjects) were used to develop the DL algorithms. The DL model was evaluated by 5-fold cross-validation and tested on an independent validation data set of 277 images. The diagnostic performance of the DL algorithm was assessed by calculating the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: Using 5-fold cross-validation during training, the average AUCs of the DL models were approximately 0.99. In the external validation data set, the DL algorithm achieved an AUC of 0.99 with a sensitivity of 94.0% and a specificity of 99.3%. The DL algorithm's performance (with an accuracy of 0.95) in diagnosing referable horizontal strabismus was better than that of the resident ophthalmologists (with accuracy ranging from 0.81 to 0.85).

Conclusions: We developed and evaluated a DL model to automatically identify referable horizontal strabismus using primary gaze photographs. The diagnostic performance of the DL model is comparable to or better than that of ophthalmologists.

Translational Relevance: DL methods that automate the detection of referable horizontal strabismus can facilitate clinical assessment and screening for children at risk of strabismus.

Introduction

Strabismus is characterized as any binocular misalignment.[1] It is traditionally considered an eye disorder affecting children, with a prevalence ranging from 0.8% to 6.0%.[2–5] In children, the most common type of strabismus is horizontal strabismus, which includes esotropia and exotropia.[6] Early detection and prompt management of strabismus in children is vital as it can improve their long-term visual and sensorimotor outcomes.[7] It has been reported that strabismus is the most common cause of childhood amblyopia.[8,9] Children with apparent strabismus may also be subject to psychosocial sequelae, such as lower self-confidence and self-esteem anxiety.[10,11]

Despite the importance of strabismus screening, there is only a limited number of pediatric ophthalmologists able to perform this task in many countries, such as China. Furthermore, clinical assessment of strabismus, such as the Hirschberg and Krimsky tests, is labor-intensive and needs to be performed by experts.[12] Choi et al. reported that even strabismus subspecialists showed underestimation or overestimation by at least 10 PD when applying the Hirschberg and Krimsky tests.[12]

Several authors have suggested automatically assessing strabismus using photographic and videographic methods.[13,14] Using a stationary photographic apparatus, Barry et al. presented a modified "self-assessment" method with satisfactory accuracy compared to orthoptic measurements of the angle of strabismus.[15] In another study, Yang et al. reported a computerized three-dimensional Strabismus Photo Analyzer, and their result was comparable with the Hirschberg and Krimsky test.[16] More recently, Lim et al. suggested a "feature-engineering" method for measuring ocular versions using photographs of the nine cardinal positions.[17] However, both manual and "feature-engineering" methods are labor-intensive and need to be specified by experts. Therefore, automated detection of strabismus, mainly referable horizontal strabismus in children, is warranted and has potential benefits in the clinical setting.

Deep learning (DL) methods, such as deep convolution neural networks (DCNNs), have demonstrated a capability to automatically screen many pediatric eye diseases, such as congenital cataract and retinopathy of prematurity, more accurately than traditional approaches.[18,19] In our previous study, we presented a DCNN that achieved automated strabismus detection based on social media photographs.[20] However, such a DL model was not developed in a clinic scenario. The screening performance of DL algorithms has not been compared with human experts using clinical assessments, such as alternate prism and cover test tests, as the gold standard.

In this study, we present DL algorithms that screen for referable horizontal strabismus on primary gaze photographs taken during a clinical assessment. We also evaluated the algorithms in an external data set to assess the performance of the DL model.

Methods

Study Design

This study followed the tenets set forth in the Declaration of Helsinki, and it was approved by the institutional review board (IRB) of both the Shanghai Children's Hospital (SCH; identifier, 2018RY029-E01) and the Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong (JSIEC; identifier, EC 20198911(4)-P11). All images were de-identified according to the Health Insurance Portability and Accountability Act Safe Harbor prior to transfer to the study investigators.[21] Informed consent was not required by the IRB due to the retrospective nature of the study and the use of fully de-identified images.

Data Sets

In this cross-sectional study, a total of 7530 primary gaze photographs (with 3330 referable horizontal strabismus and 4200 orthoptic cases) were collected from the Department of Ophthalmology, SCH (referred to as the SCH data set) from 2013 to 2019. The DL networks would require images with visually noticeable features in order to work properly. Each image was inspected by three senior pediatric ophthalmologists. If any ophthalmologist deemed an image as non-gradable (cannot visually detect strabismus), it was removed from the raw data set. We therefore excluded non-gradable images (371 [4.9%] normal and 133 [1.8%] strabismus), leaving the total data set with 7026 images (3829 normal images from 3021 orthoptic subjects and 3197 strabismus images from 2772 subjects) for DL algorithm development. The participants were children with primary horizontal strabismus who underwent surgery and orthotropic children who underwent routine refractive tests at SCH. All subjects underwent a standardized ophthalmic examination, including slit-lamp examination, Snellen visual acuity, cycloplegic refraction, and axial length with the IOLMaster (Carl Zeiss Meditec, Dublin, CA, USA). The Hirschberg test was used to assess eye alignment by noting the corneal light reflex's location within the pupil. The angle of deviation was measured by an alternate prism cover test near (0.33 m) and at a distance (6 m).

Referable horizontal strabismus was defined as follows: 1 constant infantile esotropia (≥ 40 PD), residual esotropia of accommodative esotropia (> 10 PD) with full hypermetropic correction, and intermittent or infantile exotropia (> 15 PD) with a manifestation during more than 50% of waking hours after full hypermetropic correction. Exclusion criteria included children with restrictive strabismus, sensory strabismus, paralytic strabismus, myasthenia gravis, nystagmus, and Duane syndrome. All photographs were obtained at a distance of 1 m from the subject using a commercially available camera (D800; Nikon Inc.,

Q3
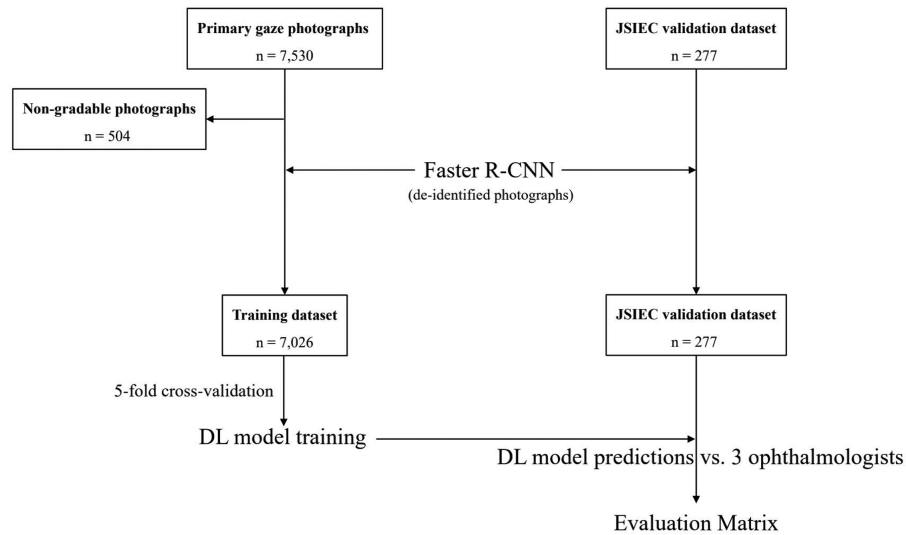
Figure 1. STARD diagrams of image datasets collection and preprocessing for detecting referable horizontal strabismus using deep learning.

Tokyo, Japan) with a pen torch attached to the camera. Only primary gaze photographs were collected for the data set.

Experimental Setup

The data set was further randomly split into a training set (80%) and a validation set (20%). We also used person-disjoint to prevent bias, meaning no subjects were shared between the training and validation data sets.[22] As our data set is not large, we used fivefold cross-validation to train and evaluate the performance of the DL model.[23] In cross-validation, the entire data set was split into five groups, and four groups were used as training data, whereas one group was used for validation. The training process was repeated 1000 times to allow for the use of all subsets exactly once as a validation data set.

We implement the DL algorithms with the TensorFlow framework (Google, version 2.1.0). The models were trained on an Ubuntu 16.04 operation system with Intel Xeon E5-2690 CPU, 128 GB RAM, and NVIDIA Titan Xp 12 GB GPU.

The algorithm's performance was measured by diagnostic parameters, including accuracy, sensitivity, specificity, and area under the receiver operating curve (AUC). The DL algorithm's performance in image classification was compared with three professional screeners on an external validation data set. All statistical analyses were carried out using the Python statistical programming language (version 3.5.1; Python Software Foundation, Beaverton, OR).

Evaluation

To evaluate the performance of the DL model outside of our in-house dataset, we also collected an external test set from the Department of Pediatrics and Strabismus, JSIEC (referred to as the JSIEC validation dataset) from 2018 to 2019 using the same study protocol as described above (Fig. 1). The JSIEC validation dataset included 277 primary gaze photographs with 133 referable horizontal strabismus and 144 normal cases.

The Architecture of the DL Algorithms

In the current study, we adopted a two-stage DL algorithm (Fig. 2) that first detects and de-identifies the primary gaze and further classifies it into referable horizontal strabismus or orthotropic. To detect regions of interest (ROI) and to satisfy the Privacy Rule's de-identification standard,[21] the ROIs (primary gaze photograph without full-face information) were first automatically localized and cropped using a faster-region based convolutional neural network (Faster R-CNN).[24] Faster R-CNN[24] is a target detection algorithm driven by the regional proposal and region-based CNN. It aims to obtain bounding boxes of the targets with location and size information. The primary gaze photographs generated by Faster-RCNN were manually checked, and necessary corrections (adjusting two eyes into a horizontal position) were made where needed. As the ROI is rectangular, we did padding to avoid changing the aspect ratio. The second stage was implemented by performing transfer
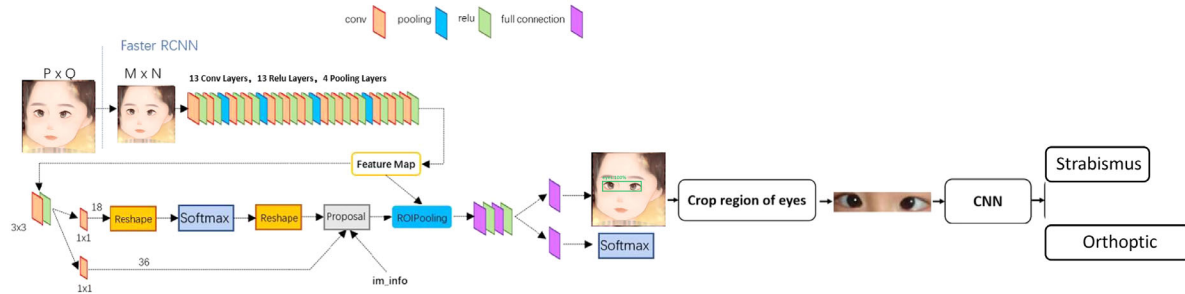
Figure 2. Diagrams showing an overview of the proposed two-stage deep learning (DL) algorithm automated detection for referable horizontal strabismus. The first stage algorithm is a faster-region based convolutional neural network (Faster R-CNN) to localize and crop the ROIs (primary gaze photograph without full-face information). The second stage algorithm is pretrained Inception-V3 networks to automatically detect referable horizontal strabismus.

learning based on the three DCNNs[25] architecture, which was pretrained on ImageNet[26] (1000 object categories with more than 1 million images). The applied DCNNs were VGG16 (Visual Geometry Group, Department of Engineering Science, University of Oxford), Inception-V3 (Google Inc.), and Xception (Google Inc.).[25–27] Image pixels were rescaled to values in a range of 0 through 1 and interpolated to fill a 299 × 299 matrix to match the pretrained networks. Data augmentation was performed to enhance the data set, including random horizontal flipping and adjustments of the saturation, brightness, and contrast. For training, we use an Adam optimizer learning rate of 0.0001 and a minibatch gradient descent of size 32. Early stopping was applied when the validation loss did not decrease for 10 epochs. To better visualize the learning procedure of our DL algorithms, we use Class Activation Maps (CAM)[28] to highlight the area that the DL algorithms may highlight in the discriminative image regions.

Comparison of the DL Algorithms with Human Experts

We used an external validation data set to compare the DL algorithm's referral decisions with the decisions made by human experts. Three resident ophthalmologists (X.L.X., H.Y., and Q.Y.) with at least 3 years of clinical experience in pediatric and strabismus, who were blinded from the data set collection procedures, were instructed to decide on each testing image independently. As mentioned previously, they were asked to identify referable horizontal strabismus by detecting misalignment of the first Purkinje images (reflection pattern from the corneal light reflex) from the center of the patient's pupil.

Results

After processing by Faster R-CNN, only ROIs with de-identified primary gaze photographs were cropped. The proportion of labels in the training and validation data sets is described in Table 1. A training course was performed before grading. The three resident ophthalmologists achieved an unweighted kappa value of 0.772 in a test set (20 normal and 20 referable horizontal strabismus photographs, respectively).

All three DL models achieved excellent performance to detect referable horizontal strabismus in primary gaze photographs. Using 5-fold cross-validation during training, the average AUCs of the DL models were 0.993 (95% confidence interval [CI] = 0.989–1.000) with the InceptionV3 model, 0.993 (95% CI = 0.989–1.000) with the VGG16 model (Supplementary Figs. 7, 8), and 0.991 (95% CI = 0.986–1.000) with the

Table 1. Summary of SCH and JSIEC Data Set for Development and Validation of Deep Learning Models

	SCH Data Set		JSIEC Data Set
	Training Dataset	Validation Dataset	External Validation Set
Referable horizontal strabismus	2558	639	133
Orthotropic	3064	765	144
Total	5622	1404	277
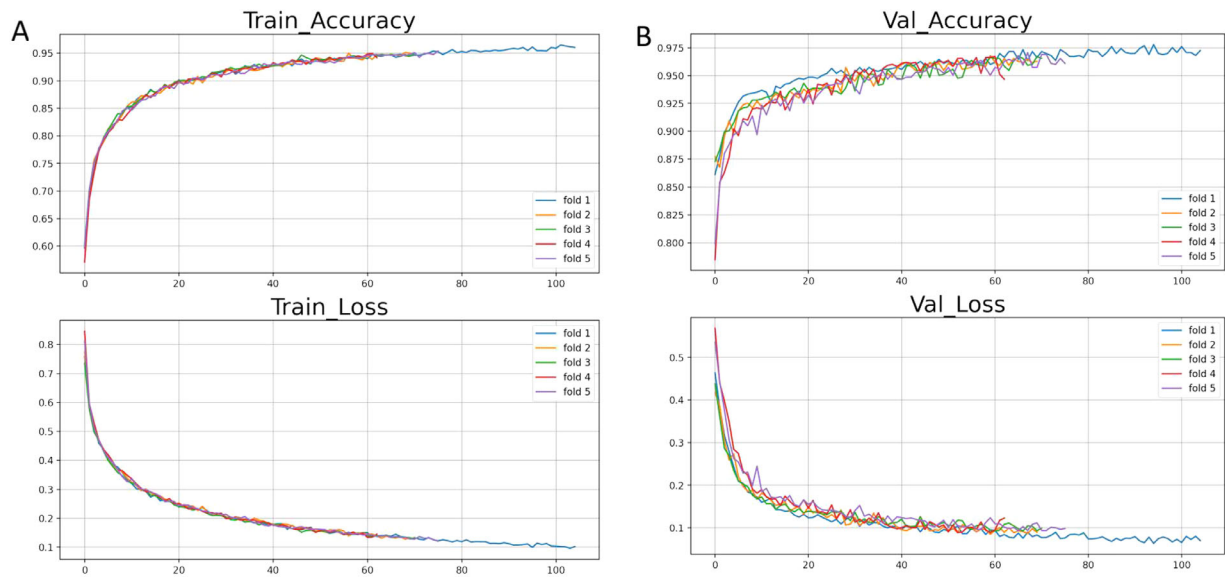
Figure 3. Plot showing the performance of the DL model. (**A**) The training accuracy and loss are plotted against epochs in the fivefold cross-validation. (**B**) Validation accuracy and loss are plotted against epochs in the fivefold cross-validation.

Xception model (Supplementary Figs. 9, 10). Considering the similar results, we only used the InceptionV3 model for further experiments in the current study.

Figures 3 and 4 show the training performance, the mean AUCs, and the confusion matrix of the DL model using 5-fold cross-validation during training. In the JSIEC validation dataset, an accuracy of 0.95 (95% CI = 0.92–0.97) with a sensitivity of 0.94 (95% CI = 0.92–0.97) and a specificity of 0.99 (95% CI = 0.98–1.00) was obtained for the classification of referable horizontal strabismus by the DL algorithm (Table 2). The diagnostic performance of the DL algorithm compared with the resident ophthalmologists is shown in Table 2 and Figure 5. The DL algorithm outperformed any ophthalmologist whose sensitivity and specificity point falls below the ROC curve of the DL algorithm (see Fig. 5).

Figures 6A and 6B presents a referable horizontal strabismus (left exotropia) photograph superimposed on its corresponding CAM (see Fig. 6B) created by the DCNN. The DL algorithms accurately identified the

left eye areas in the photographs. Figures 6C and 6D demonstrate a normal case classified by the DCNN. This child has epicanthus (Fig. 6C), which is a common cause of pseudostrabismus. Figures 6E and 6F show a failure case misclassified by the DCNNs. There are vague reflection areas in the corneal area (see Fig. 6F). The reasons for failure classification cases in all data sets are provided in Table 3. The most common reasons for failure classification were off-center (the child's eye cannot be centrally displayed due to a head tilt) and poor image quality (images with weak corneal light reflection).

Discussion

The traditional clinical assessment of strabismus, such as with the Hirschberg and Krimsky test, is labor-intensive and needs to be performed by experts. Recently, we have proposed a DL model to automate the detection of strabismus using social

Table 2. The Diagnostic Performance of DL_Model and Human Graders Testing in JSIEC Validation Data Set

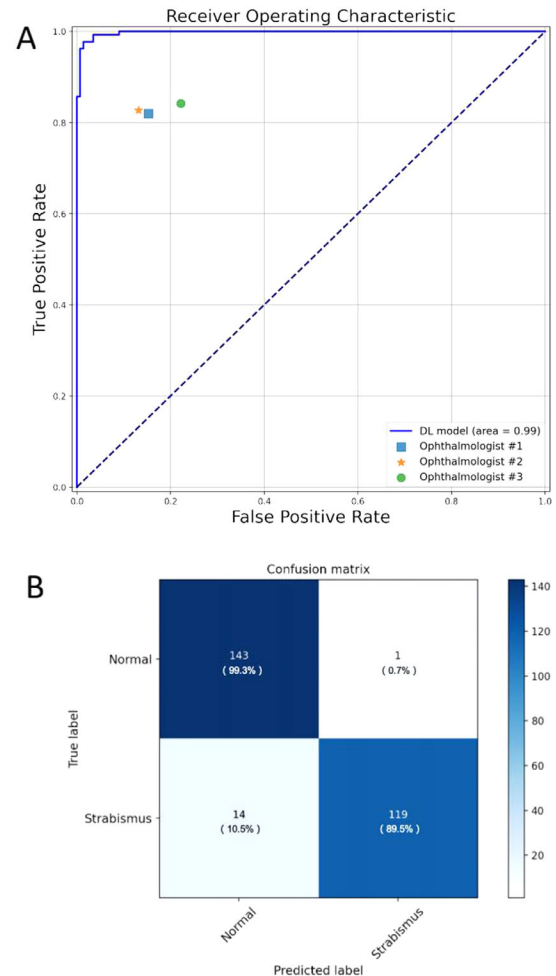|	Accuracy (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)
Deep learning models	0.968 (0.947 to 0.989)	0.993 (0.983 to 1.000)	0.940 (0.919 to 0.968)
Human graders			
Ophthalmologist #1	0.834 (0.790 to 0.878)	0.847 (0.805 to 0.889)	0.820 (0.775 to 0.865)
Ophthalmologist #2	0.848 (0.806 to 0.890)	0.868 (0.828 to 0.908)	0.827 (0.782 to 0.872)
Ophthalmologist #3	0.809 (0.763 to 0.855)	0.778 (0.729 to 0.827)	0.842 (0.799 to 0.885)

95% CI, 95% confidence interval.
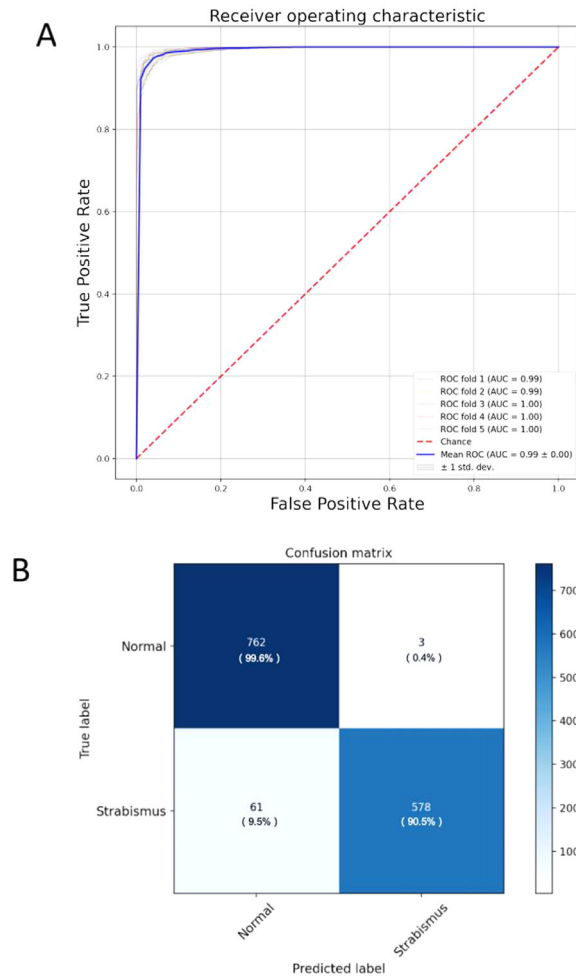
Figure 4. Mean receiver operating characteristic curves (**A**) and mean confusion matrix of the DL model for detecting referable horizontal strabismus. The performance was evaluated using fivefold cross-validation.

Figure 5. Performance of the DL model and ophthalmologists for detecting referable horizontal strabismus in the JSIEC validation dataset (**A**). Confusion matrix of the DL model in the same testing dataset (**B**).

media photographs. Despite the promising results, the previous DL model could not be substituted for ophthalmologists because the study subjects were not assessed by the Hirschberg and Krimsky test, which is the gold standard for measuring binocular alignment. In this study, using clinical evaluation as the reference, we developed and validated a DL model for automated screening of referable horizontal strabismus in children. We further validated our DL model in two different external data sets, and its performance is better than that of the human experts. To our knowledge, this is the first DL model for the screening of strabismus in children. Therefore, this study may constitute a useful baseline for applying DL to assist in detecting strabismus in the future.

Accurate, consistent assessment of binocular misalignment is particularly crucial for detecting strabismus. Compared with the previous studies, our study had a unique advantage: it allows us to train

entirely end-to-end learning instead of recognizing an output category from manual or hand-engineered features. In medical image analysis, DL has achieved excellent performance exceeding traditional diagnosis by human experts, including diabetic retinopathy, age-related macular degeneration, and glaucoma.[29–31] Our DL algorithm also demonstrated high accuracy with diagnostic performance better than that of resident ophthalmologists. Moreover, we further evaluated it using an external data set established from different hospitals. Our DL model still achieved high performance, with an AUC of 0.997 with 94% sensitivity and 99.3% specificity.

In the present study, nearly half of the misclassified cases (49.1%) were off-center (the child's eye cannot be centrally displayed due to a head tilt). The second common reason for failure of correct classification was a poor image quality (images had a weak corneal light reflex). Therefore, our method can be considered a
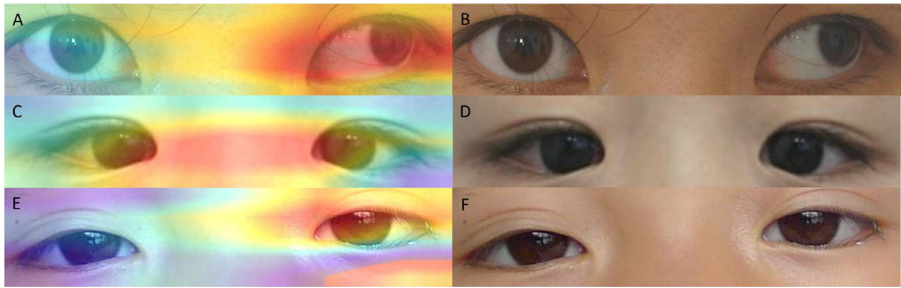
Figure 6. Examples of primary gaze photographs. (**A**) Presents a case of referable horizontal strabismus (left exotropia) superimposed with its corresponding CAM created by the DCNN. The DL algorithms accurately identified the left eye areas in the photographs. (**B**) Demonstrates a normal case classified by the DCNN. This child has epicanthus, which is a common cause of pseudostrabismus. (**C**) Shows a failure case misclassified by the DCNN. There are vague reflection areas in the corneal region.

Table 3. The Proportion of Reasons for Misclassification by the Deep Learning Model in both SCH and JSIEC Data Sets

Reason	No. (%)
Off-center of child's eye	180 (49.1)
Poor image quality	134 (36.5)
Others	53 (14.4)
Total	367 (100)

reliable tool for referable horizontal strabismus screening, which does not rely on expert input. This advantage offers our DL model an opportunity to be applied in many scenarios, such as telescreening programs, in a cost-effective and time-efficient manner.

Our study has some limitations. First, the potential for selection bias exists because most of the patients with strabismus were children who underwent surgery. Therefore, the screening performance of the DL model is less accurate in regard to variance ranges of angular deviations, such as intermittent exotropia or nystagmus. Second, our DL model was only developed among children of Chinese ethnicity; whether these results can be generalized to other populations of different ages remains to be seen. Third, the patients did not wear spectacles during photography. The angle of strabismus might be changed with spectacle-wear; thus, our DL model may not achieve good performance if glasses are prescribed. The DL networks would require images with visually noticeable features to work correctly. We excluded 504 (6.7%) of images from the raw data set, as ophthalmologists cannot visually detect strabismus from poor-quality images. With the advance of hardware and software, it is possible to design a platform to capture a good quality image and achieve better performance of a strabismus screening program.[32]

In conclusion, we demonstrated an automated DL model that can achieve high accuracy in screening referable horizontal strabismus from orthotropic based on primary gaze photographs using clinical assessment as a reference. Additional studies to determine the generalizability of the DL model, as well as its usefulness and potential cost savings in the clinical setting, seem to be warranted.

Compliance with Ethical Standards

All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments. For this type of study, formal consent is not required.

Acknowledgments

Disclosure: **C. Zheng**, None; **Q. Yao**, None; **J. Lu**, None; **X. Xie**, None; **S. Lin**, None; **Z. Wang**, None; **S. Wang**, None; **Z. Fan**, None; **T. Qiao**, None

Q6

References

1. Repka MX, Lum F, Burugapalli B. Strabismus, strabismus surgery, and reoperation rate in the united states: analysis from the IRIS Registry. *Ophthalmology*. 2018;125(10):1646–1653.

2. Robaei D, Rose KA, Kifley A, et al. Factors associated with childhood strabismus : findings from a population-based study. *Ophthalmology*. 2006;113(7):1146–1153.

3. Zhu H, Yu JJ, Yu RB, et al. Correction: Association between childhood strabismus and refractive error in Chinese preschool children. *PLoS One*. 2015;10(3):e0120720.

4. Mckean-Cowdin R, Cotter SA, Tarczyhornoch K, et al. Prevalence of amblyopia or strabismus in Asian and non-Hispanic white preschool children: multi-ethnic pediatric eye disease study. *Ophthalmology*. 2013;120(10):2117–2124.

5. Friedman DS, Repka MX, Katz J, et al. Prevalence of amblyopia and strabismus in white and African American children aged 6 through 71 months the Baltimore Pediatric Eye Disease Study. *Ophthalmology*. 2009;116(11):2128–2134.e1-2.

6. Chia A, Roy L, Seenyen L. Comitant horizontal strabismus: an Asian perspective. *Br J Ophthalmol*. 2007;91(10):1337–1340.

7. Wallace DK, Christiansen SP, Sprunger DT, et al. Esotropia and exotropia preferred practice pattern. *Ophthalmology*. 2017;125(1):143–183.

8. Mojon-Azzi SM, Kunz A, Mojon DS. Strabismus and discrimination in children: are children with strabismus invited to fewer birthday parties? *Br J Ophthalmol.* 2011;95(4):473–476.

9. Chia A, Dirani M, Chan YH, et al. Prevalence of amblyopia and strabismus in young Singaporean Chinese children. *Invest Ophthalmol Vis Sci*. 2010;51(7):3411.

10. Jackson S, Harrad RA, Morris M, Rumsey N. The psychosocial benefits of corrective surgery for adults with strabismus. *Br J Ophthalmol*. 2006;90 (7):883–888.

11. Nelson BA, Gunton KB, Lasker JN, Nelson LB, Drohan LA. The psychosocial aspects of strabismus in teenagers and adults and the impact of surgical correction. *J AAPOS*. 2008;12(1):72–76.

12. Choi RY, Kushner BJ. The accuracy of experienced strabismologists using the Hirschberg and Krimsky tests. *Ophthalmology*. 1998;105(7):1301–1306.

13. Yang HK, Han SB, Hwang JM, et al. Assessment of binocular alignment using the three-dimensional Strabismus Photo Analyzer. *Br J Ophthalmol*. 2012;96(1):78–82.

14. Effert R, Barry JC, Colberg R, et al. Self-assessment of angles of strabismus with photographic Purkinje I and IV reflection pattern evaluation. *Graefes Arch Clin Exp Ophthalmol*. 1995;233(8):494–506.

15. Barry JC, Effert R, Reim M, et al. Computational principles in Purkinje I and IV reflection pattern evaluation for the assessment of ocular alignment. *Invest Ophthalmol Vis Sci*. 1994;35(13): 4205–4218.

16. Yang HK, Seo JM, Hwang JM, et al. Automated analysis of binocular alignment using an infrared camera and selective wavelength filter. *Invest Ophthalmol Vis Sci*. 2013;54(4):2733–2737.

17. Han WL, Dong EL, Jung WL, et al. Clinical measurement of the angle of ocular movements in the nine cardinal positions of gaze. *Ophthalmology.* 2014;121(4):870–876.

18. Liu X, Jiang J, Zhang K, et al. Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. *PLoS One*. 2017;12(3):e0168606.

19. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–810.

20. Lu J, Feng J, Fan Z, et al. Automated strabismus detection based on deep neural networks for telemedicine applications. *arXiv*. 2018; abs/1809.02940. Available at: https://arxiv.org/abs/1809.02940. Accessed December 3, 2018.

21. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Health and Human Services, November 20, 2013. Office for Civil Rights. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed March 12, 2017.

22. Kuehlkamp A, Bowyer K. Predicting gender from iris texture may be harder than it seems. *arXiv*. 2018; abs/1811.10066. Available at: http://arxiv.org/abs/1811.10066. Accessed November 25, 2018.

23. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*. 1995. Available at: https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection.

24. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–1149.

25. Szegedy C, Vanhoucke V, Ioffe S, et al. (2016). Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR). 2818–2826, https://doi.org/10.1109/CVPR.2016.308.

26. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*. 2014; abs/1409.1556. Available at: https://arxiv.org/abs/1409.1556 . Accessed November 11, 2018.

27. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vision*. 2015;115, 211–252.

28. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. *arXiv*. 2015; abs/1512.04150. Available at: https://kobiso.github.iio/research/research/learning-deep-features/#:~:text=Learning%20Deep%20Features%20for%20Discriminative%20Localization%201%20Introduction.,on%20the%20ILSVRC%202014%20benchmark%20dataset.%204%20References.

29. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402.

30. Treder M, Lauermann JL, Eter N. Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(11):2053–2060.

31. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology.* 2018;125(8):1199–1206.

32. Strabismus Screening – Now There's an App for That. Roisin McGuigan, August 6, 2015. Available at: https://theophthalmologist.com/business-profession/strabismus-screening-now-theres-an-app-for-that. Accessed October 30, 2020.

translational vision science & technology