AMERICAN ACADEMY OF OPHTHALMOLOGY®

Artificial Intelligence to Differentiate Pediatric Pseudopapilledema and True Papilledema on Fundus Photographs

Melinda Y. Chang, MD,[1,2] Gena Heidary, MD, PhD,[3,4] Shannon Beres, MD,[5] Stacy L. Pineles, MD,[6] Eric D. Gaier, MD, PhD,[3,4,7] Ryan Gise, MD,[3,4] Mark Reid, PhD,[1] Kleanthis Avramidis, MEng,[8] Mohammad Rostami, PhD,[8,9] Shrikanth Narayanan, PhD,[8,9] for the Pediatric Optic Nerve Investigator Group (PONIG)

Purpose: To develop and test an artificial intelligence (AI) model to aid in differentiating pediatric pseudopapilledema from true papilledema on fundus photographs.

Design: Multicenter retrospective study.

Subjects: A total of 851 fundus photographs from 235 children (age $<$ 18 years) with pseudopapilledema and true papilledema.

Methods: Four pediatric neuro-ophthalmologists at 4 different institutions contributed fundus photographs of children with confirmed diagnoses of papilledema or pseudopapilledema. An AI model to classify fundus photographs as papilledema or pseudopapilledema was developed using a DenseNet backbone and a tribranch convolutional neural network. We performed 10-fold cross-validation and separately analyzed an external test set. The AI model's performance was compared with 2 masked human expert pediatric neuro-ophthalmologists, who performed the same classification task.

Main Outcome Measures: Accuracy, sensitivity, and specificity of the AI model compared with human experts.

Results: The area under receiver operating curve of the AI model was 0.77 for the cross-validation set and 0.81 for the external test set. The accuracy of the AI model was 70.0% for the cross-validation set and 73.9% for the external test set. The sensitivity of the AI model was 73.4% for the cross-validation set and 90.4% for the external test set. The AI model's accuracy was significantly higher than human experts on the cross validation set ($P < 0.002$), and the model's sensitivity was significantly higher on the external test set ($P = 0.0002$). The specificity of the AI model and human experts was similar (56.4%−67.3%). Moreover, the AI model was significantly more sensitive at detecting mild papilledema than human experts, whereas AI and humans performed similarly on photographs of moderate-to-severe papilledema. On review of the external test set, only 1 child (with nearly resolved pseudotumor cerebri) had both eyes with papilledema incorrectly classified as pseudopapilledema.

Conclusions: When classifying fundus photographs of pediatric papilledema and pseudopapilledema, our AI model achieved $>$ 90% sensitivity at detecting papilledema, superior to human experts. Due to the high sensitivity and low false negative rate, AI may be useful to triage children with suspected papilledema requiring work-up to evaluate for serious underlying neurologic conditions.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100496 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Ophthalmologists are frequently called upon to "rule out papilledema" or differentiate papilledema from pseudopapilledema. In children, this task is particularly difficult because optic disc drusen, the most common cause of pseudopapilledema, are frequently buried and noncalcified early in life.[1,2] Distinguishing between papilledema and pseudopapilledema is particularly important in children, since the work-up for papilledema involves neuroimaging and lumbar puncture, both of which require sedation in uncooperative patients (including most young children). Misdiagnosing pseudopapilledema as papilledema may lead to an unnecessary and invasive systemic work-up. However, misdiagnosing papilledema as pseudopapilledema may result in failure to identify and treat a serious neurologic diagnosis such as a brain tumor or meningitis. Though several investigators have published on imaging techniques

https://doi.org/10.1016/j.xops.2024.100496
ISSN 2666-9145/24

to identify papilledema and pseudopapilledema in children,[2−7] a recent literature review on this topic concluded that no single ophthalmic imaging technique accurately differentiates the 2 diagnostic entities.[8] In some cases, longitudinal follow-up may be required to clarify the diagnosis.

Given the limitations of current ophthalmic imaging techniques interpreted by humans, the Pediatric Optic Nerve Investigator Group was formed to identify new methods of differentiating pediatric papilledema from pseudopapilledema. The group's first task was to determine whether artificial intelligence (AI) techniques could be leveraged to aid in clinical decision making. The Brain and Optic Nerve Study with Artificial Intelligence (BONSAI) group has previously shown that a deep learning model can achieve high accuracy in classifying fundus photographs as papilledema versus normal or nonpapilledema optic nerve abnormality.[9] However, the BONSAI model primarily included data from adults and the comparison group included a heterogeneous set of diagnoses, including normal optic nerves.

The purpose of the present study is to report the accuracy of a deep learning model in classifying fundus photographs as pediatric papilledema or pseudopapilledema. We trained, validated, and externally tested a deep learning model using fundus photographs submitted from 4 separate institutions. For comparison, the same set of fundus photographs was reviewed and classified by 2 masked expert pediatric neuro-ophthalmologists.

Methods

We performed a retrospective multicenter review of fundus photographs of children with papilledema and pseudopapilledema submitted by 4 pediatric neuro-ophthalmologists (M.Y.C., S.L.P., G.H., and S.B.). This study was approved by the institutional review board at each separate institution and adhered to the tenets of the Declaration of Helsinki and the United States Health Insurance Portability and Accountability Act of 1996. Informed consent was waived by the institutional review board for this retrospective study. Children with papilledema and pseudopapilledema were identified either through an electronic medical records search using International Classification of Diseases (ICD) codes (papilledema: ICD-9377.0, 348.0, ICD-10 H47.11, G93.2; pseudopapilledema: ICD-9377.24, 377.21, ICD-10 H47.33, H47.32) or by searching an institutional database of patients with these diagnoses. Charts were reviewed by the submitting neuro-ophthalmologist to determine if patients met inclusion criteria for the study: age < 18 years at the time of fundus photograph acquisition, in addition to the following diagnostic criteria: children with papilledema were required to have elevated optic nerves on fundoscopy and/or OCT as well as neuroimaging evidence of intracranial pathology causing increased intracranial pressure (such as a brain tumor), or a lumbar puncture with opening pressure > 28 cm H_2O. Children with pseudopapilledema were required to have either normal neuroimaging (with normal brain parenchyma, no signs of hydrocephalus, mass, or structural lesion, and no meningeal enhancement) and lumbar puncture with normal opening pressure (< 28 cm H_2O), or longitudinal follow-up for ≥ 6 months demonstrating no change in optic nerve appearance. Longitudinal follow-up was required even in cases where optic disc drusen were demonstrated on ophthalmic imaging. We excluded cases with papilledema superimposed on pseudopapilledema. We also excluded patients with resolved papilledema/gliosis and optic atrophy, as well as fundus photographs that were taken using wide-field fundus cameras and photographs deemed by the submitting pediatric neuro-ophthalmologist to be too poor in quality for accurate classification.

Data Collection

After identifying patients with clinical characteristics and fundus photographs that met inclusion and exclusion criteria, the pediatric neuro-ophthalmologists collected all fundus photographs meeting criteria for the study and uploaded them to a Health Insurance Portability and Accountability Act-compliant REDCap electronic data capture tool hosted at the University of Southern California.[10] We included photographs of both eyes over multiple visits if available. Photographs from subsequent visits were only included if active papilledema was seen on fundoscopy and/or OCT. As noted above, fundus photographs of resolved papilledema, gliosis, and optic atrophy were excluded. Additional data entered into the REDCap database included study site, patient demographics, clinical characteristics (diagnosis, cause of papilledema or pseudopapilledema, neuroimaging results, and lumbar puncture results), duration of follow-up, and dates and number of visits with fundus photographs. Data were deidentified and patients were assigned a subject identification number by REDCap.

Masked Classification of Fundus Photographs by Human Experts

Two pediatric neuro-ophthalmologists who did not contribute fundus photographs for this study were recruited to serve as masked experts (E.D.G. and R.G.). Fundus photographs were randomized in order and renamed in a deidentified manner, so that photographs from the same patient (different eye or visit) would not be adjacent to one another in the list of photographs. The masked experts independently classified each photograph as papilledema or pseudopapilledema, without any rules provided to guide their decision-making. The performance of masked experts was compared with the AI model (see Statistical Analysis, below).

Consensus Grading of Papilledema Severity

In order to determine whether the accuracy of the AI model depended on the severity of papilledema, the 4 submitting pediatric neuro-ophthalmologists performed Frisen grading of the papilledema photographs. The original Frisen grading paper was reviewed prior to independent grading.[11] The fundus photographs that received identical scores on ≥ 3 out of 4 grades were considered to have reached consensus (the majority score was used as the grade). The remaining fundus photographs were reviewed during a group meeting with all 4 pediatric neuro-ophthalmologists. After discussion, a consensus grade was agreed upon for all photographs.

Development of AI Model

The algorithm used to differentiate between papilledema and pseudopapilledema cases included 2 main components—an unsupervised optic disc detector and a deep neural network for the classification task. The disc detector takes a raw fundus image as input, locates the optic disc through morphological operations, and outputs a cropped image of the disc.[12] The algorithm utilizes the fact that the optic disc is one of the brightest elements in a fundus image and applies intensity thresholding. To eliminate false positives, a detected region is selected as a candidate optic disc only after evaluating its eccentricity and covered area. Since this is an unsupervised approach, a small subset of images that the detector failed (< 15%) were manually processed and cropped.

For the classification task, we utilized a DenseNet backbone model, already pretrained on the ImageNet dataset for general image classification and applied a multibranch training method.[12] According to this approach, 3 different DenseNet instances were created and trained on the same data, under different color transformations of the cropped fundus. Specifically, we randomly modified the contrast intensity of the red and the green channel to create 2 additional input views for training. The additive contrast ranged between 20% and 60% and was randomly selected for each training sample at each step. These transformations on the color channels of the fundus were selected to unveil particular biomarkers, such as hemorrhages and vessel congestion. We note that we did not apply any transformation during the test phase. The 3 output feature vectors were averaged and a fully connected layer was used to generate the class predictions. We trained for a maximum of 50 epochs in a 10-fold cross-validation regime among the available subjects, from which we recorded, for all available images, the test probability of papilledema. Since this regime allowed for more than a single prediction for each image we finally considered the average output probabilities.

We then created an external test set using a random 20% sample of the fundus photographs (selected by patient, such that there were no patients with photographs in both the external test and training/validation sets), matched to the original sample on proportion of photographs with each diagnosis from each clinical site. The remaining 80% of fundus photographs were used for training and validation and the papilledema probability was calculated in each of the photos in the previously unseen external test set.

Class Activation Maps

To visualize the key optic disc features that contributed to the neural network's performance, we utilized class activation maps extracted from the model gradients. This involves taking an already trained model and a test image, computing the model's predictions, and tracing the image gradients back to the input layer. This process highlights the areas of the image whose parameters in the input layer have the highest values and that the model prioritizes in its prediction outcome.

Statistical Analysis

A receiver operating curve for the AI model was generated, and the area under the receiver operating curve (AUC) was calculated. The AI model and the human experts were compared with the gold standard of clinical diagnosis by the submitting pediatric neuro-ophthalmologist. The clinical diagnosis was based on history, examination, ancillary ophthalmic imaging tests ordered at the clinician's discretion (including ultrasonography, OCT, autofluorescence, and fluorescein angiography), neuroimaging and lumbar puncture results if performed, and longitudinal follow-up. The accuracy, sensitivity, and specificity of the AI model varied based on the papilledema probability cut-off value that was used to classify photographs as pseudopapilledema or papilledema. In order to maximize sensitivity while maintaining specificity similar to human experts, a cut-off value of 0.45 was chosen. Using this threshold, the accuracy, sensitivity, specificity, and positive and negative predictive values of the overall model (both the cross-validation and external test sets) were calculated. These metrics were also used to evaluate limited models including (1) grade 1 to 2 papilledema plus all pseudopapilledema fundus photographs (model for mild papilledema); (2) grade 3 to 5 papilledema plus all pseudopapilledema fundus photographs (model for moderate-to-severe papilledema); and (3) only the first set of fundus photographs per patient (model for initial presentation).

Masked expert fundus photograph classification was evaluated by the same metrics. Sensitivity, specificity, and overall accuracy were compared between each expert and the AI model using McNemar's tests. Positive and negative predictive values were compared between AI and human experts using the Generalized Score Statistic.[13] STATA (version 15.1; Stata Corp) was used for statistical analyses. The P value threshold for significance was set to 0.000625 using a Bonferroni adjustment for 60 comparisons and familywise α of 0.05.

Results

We included 851 fundus photographs of 235 children (105 with papilledema and 130 with pseudopapilledema). The demographics of included patients are shown in Table 1. The causes of papilledema and pseudopapilledema are provided in Table S1. The mean age at diagnosis was older in patients with papilledema (11.7 vs. 10.2 years, $P = 0.0028$). Females predominated in both groups, with no difference in sex distribution between groups. There was no difference in race or ethnicity between groups. Overall, the majority of patients were White, but Hispanic patients represented a significant minority (31%). There was a significant difference in distribution of patients by site, with University of California, Los Angeles and Stanford contributing relatively fewer patients with papilledema than the other sites. The most common cause of papilledema was pseudotumor cerebri syndrome (78%). The most common cause of pseudopapilledema was optic disc drusen (61%). Among patients with papilledema, 83 (79%) had lumbar puncture results available (in others, lumbar puncture was contraindicated due to a space-occupying lesion in the brain); the average opening pressure was 40 ± 10 cm H_2O.

Of the 851 fundus photographs, 380 were from children with papilledema and 471 were from children with pseudopapilledema. The distribution of fundus cameras used to acquire these photographs by site is shown in Table S2. Different fundus cameras were used at each site to acquire mydriatic or nonmydriatic, nonwidefield fundus photographs.

The receiver operating characteristic curve and accuracy, sensitivity, and specificity of the cross-validation model with various papilledema probability cut-off values are shown in Figure 1. The AUC on the cross-validation set was 0.77 (95% confidence interval 0.74−0.80). With a papilledema probability cut-off value of 0.45 (ie, probabilities $\geq$ 0.45 were classified as papilledema, and probabilities < 0.45 were classified as pseudopapilledema), the accuracy, sensitivity, and specificity were 70.0%, 73.4%, and 67.3%, respectively (Table 2). The AI model achieved significantly higher accuracy and sensitivity than expert 2 (70.0% vs. 61.0% and 73.4% vs. 62.9%, $P < 0.0001$ and $P = 0.0003$, respectively) on the cross-validation set. Positive and negative predictive values were also significantly higher than expert 2 (64.4% vs. 55.6% and 75.8% vs. 66.5%, respectively; $P < 0.0001$ for both comparisons). Specificity was not significantly different from experts 1 and

Table 1. Demographics of 235 Children With Papilledema and Pseudopapilledema Included in This Study

	Papilledema n = 105	Pseudopapilledema n = 130	*P* Value
Age (mean ± SD) at diagnosis	11.7 ± 4.0 years	10.2 ± 3.5	0.0028
Sex (female/male)	71 (68%)/34 (32%)	175 (76%)/55 (24%)	0.10
Race			0.11
Asian	2 (2%)	8 (6%)	
Black or African American	5 (5%)	2 (1.5%)	
Native Hawaiian or Pacific Islander	1 (1%)	0	
White	66 (63%)	83 (64%)	
Other	21 (20%)	17 (13%)	
Decline	10 (10%)	20 (15%)	
Ethnicity			0.07
Hispanic	37 (35%)	36 (28%)	
Non-Hispanic	53 (50%)	60 (46%)	
Decline	15 (14%)	34 (26%)	
Site			< 0.0001
Boston Children's Hospital	42 (40%)	30 (24%)	
Children's Hospital Los Angeles	38 (36%)	25 (20%)	
Stanford University	19 (18%)	38 (29%)	
UCLA	6 (6%)	34 (27%)	

SD = standard deviation; UCLA = University of California, Los Angeles.

2 after correcting for multiple comparisons (67.3% vs. 60.3% and 59.4%, $P = 0.02$ and $P = 0.008$, respectively).

The performance of the AI model on the external test set is also provided in Table 2. The model achieved higher sensitivity, negative predictive value, and overall accuracy in the external test set compared with the cross-validation set. The AUC for the external test set was 0.81 (95% confidence interval 0.74−0.87). The accuracy, sensitivity, and specificity were 73.9%, 90.4%, and 56.4%, respectively. On external testing, the AI model was significantly more sensitive at detecting papilledema than both human experts (90.4% vs. 68.7%, $P = 0.0002$). Specificity did not significantly differ between the AI model and expert 1 or 2 on the external test set (56.4% vs. 61.5% and 65.4%, $P = 0.49$ and $P = 0.27$, respectively). Negative predictive value was higher in the AI model compared with both experts on the external test set (84.6% vs. 64.9% [expert 1] and 66.2% [expert 2]), although this difference did not reach statistical significance after correcting for multiple comparisons.

In the external test set, there were only 8 false negatives (i.e., fundus photographs of papilledema incorrectly classified as pseudopapilledema). Of these, 6 had correct classification of the fellow eye as papilledema. The single patient with papilledema whose fundus photographs of both eyes were misclassified as pseudopapilledema was a 16-year-old girl with pseudotumor cerebri syndrome and grade 1, nearly resolved papilledema (Figure 2). Thus, the model did not misclassify any patients with sight-threatening papilledema or papilledema from a life-threatening etiology.

To determine the effect of papilledema grade, performance metrics of the AI model and expert graders were repeated in subgroup analyses of photos demonstrating mild and moderate-to-severe papilledema. The distribution of papilledema grades is shown in Table S3. Mild papilledema (grade 1−2) was demonstrated in 204 (54%) photographs, and moderate-to-severe papilledema (grade 3−5) was seen in 176 (46%) photographs. Table 3 shows the accuracy, sensitivity, specificity, and positive and negative predictive values of the AI model and human experts in classifying fundus photographs of mild and moderate-to-severe papilledema. The sensitivity of the AI model in detecting mild papilledema was 64.2% in the cross-validation set, higher than both human experts (52.0% and 38.7%, $P = 0.006$ and $P < 0.0001$, respectively). The AI model's sensitivity on the external test set was 87.8%, significantly higher than both human experts (53.1% and 49.0%, $P = 0.0001$ and $P < 0.0001$, respectively). Both positive and negative predictive value were higher in the AI model than both human experts in the cross-validation set, and the AI model also

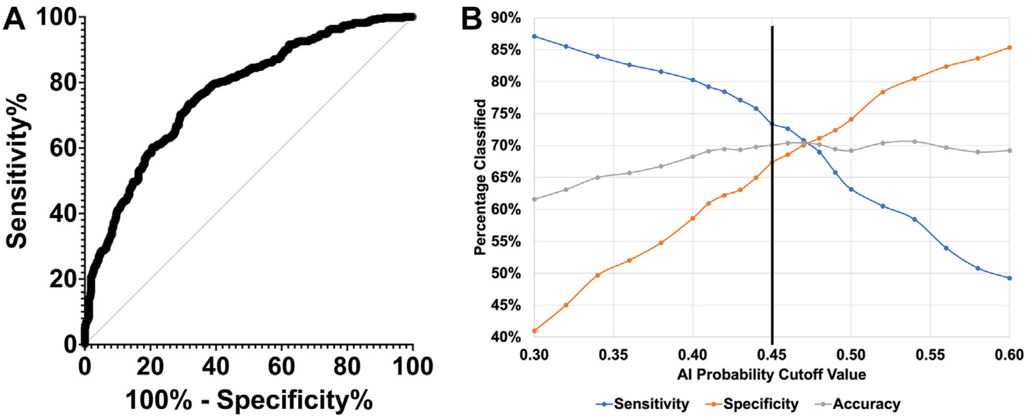

Figure 1. A, Receiver operating characteristic curve of the artificial intelligence (AI) model to classify pediatric papilledema and pseudopapilledema fundus photographs on the cross-validation set. **B,** Accuracy, sensitivity, and specificity of the AI model at varying papilledema probability cut-off values. A cut-off of 0.45 was chosen for this study to maximize sensitivity while maintaining specificity similar to human experts.

Table 2. AUC, Accuracy, Sensitivity, Specificity, and Positive and Negative Predictive Values of the AI Model, Expert Grader 1, and Expert Grader 2 in Classifying Fundus Photographs of Pediatric Papilledema and Pseudopapilledema

	AI Model	Expert 1 (E1)	Expert 2 (E2)	*P* Value AI vs. E1	*P* Value AI vs. E2
Cross-validation set					
AUC	0.77 (0.74–0.80)				
Accuracy	70.0 (66.8–73.1)	63.6 (60.2–66.8)	61.0 (57.6–64.3)	0.002	< 0.0001*
Sensitivity	73.4 (68.7–77.8)	67.6 (62.7–72.3)	62.9 (57.8–67.8)	0.04	0.0003*
Specificity	67.3 (62.9–71.5)	60.3 (55.7–64.7)	59.4 (54.9–63.9)	0.02	0.008
Positive predictive value	64.4 (59.7–68.9)	57.9 (53.1–62.5)	55.6 (50.7–60.3)	0.003	< 0.0001*
Negative predictive value	75.8 (71.4–79.9)	69.8 (65.1–74.2)	66.5 (61.8–71.0)	0.005	< 0.0001*
External test set					
AUC	0.81 (0.74–0.87)				
Accuracy	73.9 (66.4–80.5)	65.2 (57.3–72.5)	67.1 (59.2–74.3)	0.07	0.17
Sensitivity	90.4 (81.9–95.7)	68.7 (57.6–78.4)	68.7 (57.6–78.4)	0.0002*	0.0002*
Specificity	56.4 (44.7–67.6)	61.5 (49.8–72.3)	65.4 (53.8–75.8)	0.49	0.27
Positive predictive value	68.8 (59.2–77.3)	65.5 (55.1–75.4)	67.9 (56.8–77.6)	0.45	0.85
Negative predictive value	84.6 (71.9–93.1)	64.9 (52.9–75.6)	66.2 (54.6–76.6)	0.0008	0.002

AI = artificial intelligence; AUC = area under the receiver operator characteristic curve.
Results from cross-validation and external test sets are displayed separately. Percentages are shown with 95% confidence intervals in parenthesis. *P* values comparing the AI model with expert graders are provided. Significant values after correcting for multiple comparisons (< 0.000625) are indicated by asterisks.

had significantly higher negative predictive value than both experts on the external test set for mild papilledema (88% vs. 67.6% and 67.1%, respectively; $P = 0.0002$ for both comparisons). The AI model and human experts had similar specificity when classifying mild papilledema. Additionally, the AI model and human experts performed similarly in accuracy and sensitivity when classifying moderate-to-severe papilledema. On the moderate-to-severe cases in the external test set, specificity was lower in the AI model compared with human experts (56.4% vs. 61.5% and 65.4%), but this difference did not reach statistical significance ($P = 0.49$ and $P = 0.27$).

A final subanalysis was performed including only the fundus photographs from each patient's first ophthalmology visit (Table S4). The AI model (on both cross-validation and external test sets) had higher sensitivity, negative predictive value, and accuracy on this subgroup analysis compared with human experts, although differences did not reach statistical significance with the Bonferroni correction.

Finally, class activation maps highlighting the regions of each fundus photograph used by the AI model for classification were generated. Figure 3 displays an example of a fundus photograph of a child with pseudotumor cerebri that was incorrectly classified as pseudopapilledema by the AI model but correctly identified as papilledema by both human experts. Peripapillary hemorrhages, a sign of true papilledema, were cropped by the automated algorithm for optic nerve detection (Fig 3B). The class activation map (Fig 3C) indicates that the model focused primarily on the temporal optic nerve to make the incorrect classification decision, rather than the optic nerve borders or peripapillary region.

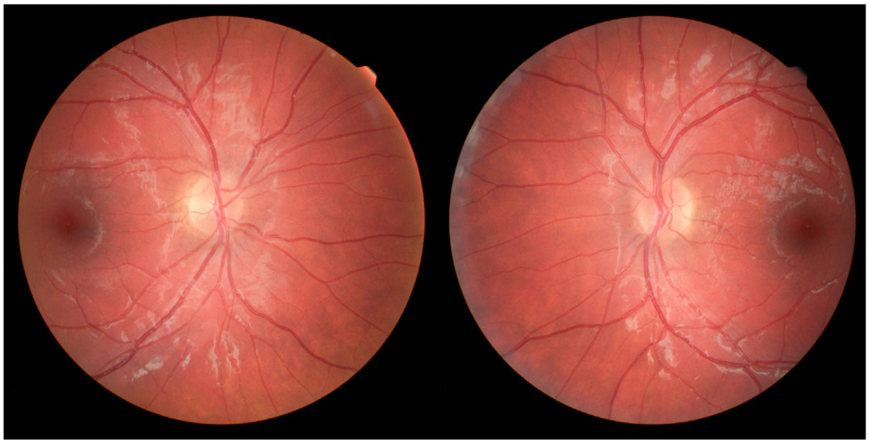

Figure 2. Fundus photographs of the single patient with papilledema with misclassification of both eyes by the artificial intelligence model in the external testing set.

Table 3. AUC, Accuracy, Sensitivity, Specificity, and Positive and Negative Predictive Values of the AI Model, Expert Grader 1, and Expert Grader 2 in Classifying Fundus Photographs of Pediatric Papilledema and Pseudopapilledema, by Papilledema Grade

	AI Model	Expert 1 (E1)	Expert 2 (E2)	*P* Value AI vs. E1	*P* Value AI vs. E2
Mild papilledema (Frisen grade 1–2)					
Cross validation set					
AUC	0.71 (0.67–0.74)				
Accuracy	66.4 (62.7–69.9)	57.8 (54.0–61.5)	53.2 (49.3–57.0)	0.0006*	< 0.0001*
Sensitivity	64.2 (57.2–70.8)	52.0 (44.9–59.0)	38.7 (32.0–45.8)	0.006	< 0.0001*
Specificity	67.3 (62.9–71.5)	60.3 (55.7–64.7)	59.4 (54.9–63.9)	0.02	0.008
Positive predictive value	46.0 (40.1–51.9)	36.2 (30.7–42.0)	29.3 (23.9–35.1)	0.0003*	< 0.0001*
Negative predictive value	81.3 (77.1–85.0)	74.3 (69.7–78.7)	69.1 (64.4–73.6)	0.0005*	< 0.0001*
External test set					
AUC	0.78 (0.70–0.85)				
Accuracy	68.5 (59.7–76.5)	58.3 (49.2–67.0)	59.1 (50.0–67.7)	0.07	0.13
Sensitivity	87.8 (75.2–95.4)	53.1 (38.3–67.5)	49.0 (34.4–63.7)	0.0001*	< 0.0001*
Specificity	56.4 (44.7–67.6)	61.5 (49.8–72.3)	65.4 (53.8–75.8)	0.49	0.27
Positive predictive value	55.8 (44.1–67.2)	46.4 (33.0–60.3)	47.1 (32.9–61.5)	0.09	0.17
Negative predictive value	88.0 (75.7–95.5)	67.6 (55.5–78.2)	67.1 (55.4–77.5)	0.0002*	0.0002*
Moderate to severe papilledema (Frisen grade 3–5)					
Cross validation set					
AUC	0.84 (0.81–0.87)				
Accuracy	71.9 (68.2–75.3)	67.2 (63.5–70.8)	68.0 (64.2–71.6)	0.05	0.10
Sensitivity	84.1 (77.8–89.2)	85.8 (79.7–90.6)	90.9 (85.7–94.7)	0.62	0.03
Specificity	67.3 (62.9–71.5)	60.3 (55.7–64.7)	59.4 (54.9–63.9)	0.02	0.008
Positive predictive value	49.0 (43.2–54.8)	44.7 (39.3–50.1)	45.6 (40.3–51.0)	0.06	0.12
Negative predictive value	91.9 (88.5–94.5)	91.9 (88.3–94.7)	94.6 (91.4–96.9)	0.99	0.09
External test set					
AUC	0.84 (0.76–0.90)				
Accuracy	67.8 (58.4–76.4)	70.5 (61.2–78.8)	75.0 (65.9–82.7)	0.63	0.23
Sensitivity	94.1 (80.3–99.3)	91.2 (76.3–98.1)	97.1 (84.7–99.9)	0.65	0.56
Specificity	56.4 (44.7–67.6)	61.5 (49.8–72.3)	65.4 (53.8–75.8)	0.49	0.27
Positive predictive value	48.5 (36.0–61.1)	50.8 (37.7–63.9)	55.0 (41.6–67.9)	0.63	0.23
Negative predictive value	95.7 (85.2–99.5)	94.1 (83.8–98.8)	98.1 (89.7–100)	0.73	0.49

AI = artificial intelligence; AUC = area under the receiver operator characteristic curve.
Results from cross-validation and external test sets are displayed separately. Percentages are shown with 95% confidence intervals in parenthesis. *P* values comparing the AI model with expert graders are provided. Significant values after correcting for multiple comparisons (< 0.000625) are indicated by asterisks.

Discussion

In this study, we developed a deep learning AI model to differentiate pediatric papilledema from pseudopapilledema using transfer learning and a tribranch convolutional neural network. On external testing, the overall model achieved an AUC of 0.81. With the threshold optimized to maximize sensitivity while maintaining specificity similar to human experts, the accuracy, sensitivity, and specificity of the model on external testing were 73.9%, 90.4%, and 56.4%, respectively. The model was significantly more sensitive at detecting papilledema than human experts, and the difference between AI and human experts was especially marked in cases of mild (grade 1–2) papilledema. We chose to optimize sensitivity given that the consequences of misdiagnosing papilledema (potentially missing a life-threatening diagnosis such as a brain tumor) are more serious than mistaking pseudopapilledema for papilledema (resulting in unnecessary diagnostic work-up including neuroimaging and/or lumbar puncture). Importantly, in the external test set of 161 photographs, there were only 8 fundus photographs of papilledema that were misclassified. When these misclassifications were evaluated on a patient basis, there was only 1 patient with papilledema whose fundus photographs from both eyes were misclassified. This patient had nearly resolved papilledema from pseudotumor cerebri. Thus, on external testing, our model did not misclassify any patients with life- or vision-threatening papilledema.

The accuracy of our human graders (61.0%–67.1%) is similar to previous reports in the literature. In 2017, Chang et al[2] conducted a prospective study in which 3 masked expert neuro-ophthalmologists graded fundus photographs (and other imaging modalities) as representing pediatric pseudopapilledema or papilledema. The accuracy of classifying fundus photographs ranged from 63% to 71% for the 3 individual graders. The authors concluded that no imaging modality interpreted in isolation, including fundus photographs, achieved sufficient accuracy for differentiating pediatric papilledema from pseudopapilledema.

In contrast, the BONSAI group developed an AI model that detected papilledema on fundus photographs with an AUC of 0.96 and accuracy of 87.5% in the external testing data set.[9] Major differences between the current study and

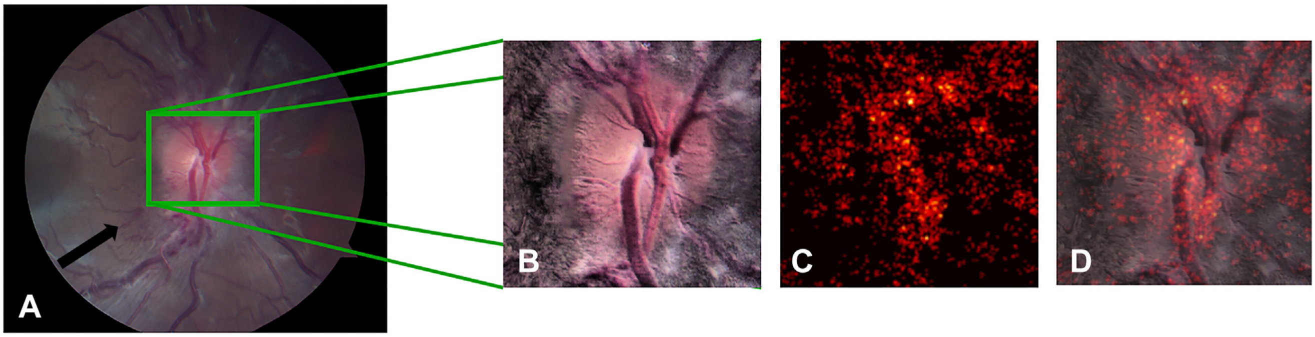

Figure 3. Fundus photograph of a child with papilledema classified incorrectly by the artificial intelligence model and correctly by both human experts. The full photograph (**A**) demonstrating peripapillary hemorrhages (arrow) was cropped close to the optic nerve (**B**) using the unsupervised algorithm for optic nerve detection. The class activation map (**C**) highlights the areas of the photograph used by the model to incorrectly classify as pseudopapilledema. The superimposed cropped photograph and class activation map are shown in (**D**).

the BONSAI study include (1) the average age of patients in the BONSAI study was 48.6 years, whereas our study included only pediatric patients; and (2) the BONSAI model was trained to differentiate photographs of papilledema from normal optic nerves and optic nerves with any other abnormality (including optic atrophy and congenital optic disc anomalies). Pediatric pseudopapilledema, as indicated by its name, is more similar in appearance to papilledema than normal optic nerves and other optic nerve pathologies. Therefore, the classification task for our model was more difficult than the BONSAI study and lower accuracies are expected. The difference in task difficulty is reflected in the difference in performance of human experts in our study compared with the BONSAI study (accuracy 61.0%−67.1% in our study vs. 89.0%−89.6% in the BONSAI study).[14] Other smaller studies reported similarly high accuracy when using AI to detect papilledema on fundus photographs,[15−17] but these studies also used variable comparison groups (including normal optic nerves) and did not report patient ages.

The most challenging clinical scenario in evaluating children with apparently swollen optic nerves is differentiating mild papilledema from pseudopapilledema. The AI model proved to be particularly helpful in this situation, as the sensitivity for detecting mild papilledema was 87.8% on external testing, clinically and statistically significantly superior to human experts (53.1% and 49.0%, $P \leq 0.0001$). Both the AI model and human experts achieved high sensitivity in detecting moderate-to-severe papilledema (94.1 vs. 91.2% and 97.1%, $P = 0.65$ and $P = 0.56$, respectively).

Despite the AI model's superiority to human experts, the accuracy and sensitivity are still insufficient for its use as the sole factor in deciding whether to pursue work-up for papilledema. Because we used a deep neural network, the image characteristics used by the model for classification are largely unknown. However, 1 possible factor contributing to reduced model accuracy may be postulated from the class activation map in Figure 3A. This fundus photograph of papilledema was incorrectly classified as pseudopapilledema and the salient features for human interpretation (peripapillary hemorrhage and obscuration of blood vessels) were missed by the model due to excessively tight cropping. These findings suggest that our model may benefit from adjusting the cropping area after the optic nerve is detected by the unsupervised algorithm.

Given that our model was particularly sensitive in detecting mild papilledema, we propose that AI may be useful in the future to triage children with apparently mildly swollen optic nerves prior to evaluation by pediatric neuro-ophthalmology. Those with a high probability of true papilledema may be referred for immediate assessment in the emergency department, whereas those with a lower probability may be considered for urgent outpatient evaluation. Moderate-to-severe papilledema is likely to be accurately detected by both humans and AI, although not all providers may reach the level of expertise of our graders.

Our study highlights the need for additional research on methods to improve our ability to differentiate pediatric papilledema from pseudopapilledema. The next step for the Pediatric Optic Nerve Investigator Group is to initiate a multicenter, multimodality prospective study to evaluate whether a combination of imaging techniques (ultrasonography, autofluorescence, OCT, fundus photography, and fluorescence angiography) and AI can accurately distinguish between the 2 diagnoses.

The strengths of this study include a relatively diverse data set with regards to ethnicity (31% Hispanic), acquisition of fundus photographs using multiple types of mydriatic and nonmydriatic fundus cameras and multiple different operators, and strict inclusion criteria minimizing the possibility of mislabeling the "gold standard" clinical diagnosis. All patients with papilledema had abnormal neuroimaging or elevated lumbar puncture opening pressure and all patients with pseudopapilledema either had normal neuroimaging and lumbar puncture opening pressure or no change in optic nerve appearance over at least 6 months. While the latter criterion may not absolutely exclude true papilledema, the average follow-up time in the pseudopapilledema patients was 28 months. After 2 years, some degree of atrophy or gliosis would be expected if these patients had true papilledema.

The limitations of this study include the retrospective nature and relatively small sample size. An automated method of detecting the optic nerve region was developed and used for this study, but this algorithm failed in a small subset (< 15%) of photographs, which required manual cropping. Refinement of the optic nerve detection algorithm is needed prior to AI model deployment for clinical purposes. Finally, the fundus photographs were submitted by pediatric neuro-ophthalmologists at tertiary care academic centers and may not be representative of the distribution and underlying causes of pediatric papilledema and pseudopapilledema in the general population. Specifically, the number of patients with papilledema was nearly equal to those with pseudopapilledema in our cohort, whereas pseudopapilledema is far more common (up to 10 times more frequent) in the general pediatric population.[18] Although the positive predictive value of the AI model in our external test set was 69%, the positive predictive value would be much lower (18.5%) if the sensitivity and specificity of our model was applied to the general population with an estimated 10% prevalence of papilledema among children referred for elevated optic nerves. The skewed distribution is likely related to the recruitment sites, which were all highly subspecialized pediatric neuro-ophthalmic clinics at tertiary care referral centers, as well as a possible bias toward taking fundus photographs in patients with papilledema in order to monitor for changes. Future iterations of this AI model will benefit from inclusion of images from general pediatric ophthalmology clinics with a distribution of papilledema and pseudopapilledema photographs that is more representative of the general population.

In conclusion, we report that an AI model can detect papilledema with higher accuracy and sensitivity than human experts and may be particularly helpful in differentiating mild papilledema from pseudopapilledema. However, further studies with additional data, ideally including a more representative ratio of papilledema to pseudopapilledema images, are needed. Larger datasets will improve the accuracy of the AI model beyond this proof-of-concept, pilot study. A future multicenter, multimodality prospective study incorporating AI may further advance our ability to differentiate between pediatric papilledema and pseudopapilledema, reducing the likelihood of missing space-occupying lesions in the brain or subjecting children to unnecessary and invasive procedures.

Acknowledgments

The authors thank members of the Pediatric Optic Nerve Investigator Group (PONIG) who contributed to this work but did not meet authorship criteria: Mark Borchert, Rasika Sudharshan, and Mariella Saludares.

Footnotes and Disclosures

Originally received: June 19, 2023.
Final revision: February 14, 2024.
Accepted: February 15, 2024.
Available online: February 20, 2024. Manuscript no. XOPS-D-23-00120.

[1] Division of Ophthalmology, Children's Hospital Los Angeles, Los Angeles, California.

[2] Roski Eye Institute, Keck School of Medicine, University of Southern California, Los Angeles, California.

[3] Department of Ophthalmology, Boston Children's Hospital, Boston, Massachusetts.

[4] Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts.

[5] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, California.

[6] Department of Ophthalmology, Stein Eye Institute, University of California, Los Angeles, Los Angeles, California.

[7] Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, Massachusetts.

[8] Viterbi School of Engineering, University of Southern California, Los Angeles, California.

[9] Information Sciences Institute, University of Southern California, Los Angeles, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

M.Y.C.: Grants — NIH NEI K23EY033790, Knights Templar Eye Foundation, The Saban Research Institute, Children's Eye Foundation of AAPOS, Research to Prevent Blindness, and Blind Children's Center; Patents — Patent application # 63249727.

K.A.: Financial support — Academic research time as PhD student at USC.

E.D.G.: Financial support — NIH K08 EY030164; Grants — Alcon Research Institute; Royalties and Patent — Luminopia Inc; Consultant and Scientific Advisory Board — Stoke Therapeutics.

G.H.: Grants — NIH R01EY029687; Honoraria — $1000 UCSF Guest Lecturer Honorarium.

S.N.: Financial support — Academic research time as faculty at USC.

S.L.P.: Financial support — Research to Prevent Blindness, Unrestricted Funds, National eye institute; Leadership — Unpaid board member, Horizon Surgical.

M.W.R.: Financial support — Larry & Celia Moh Foundation and private donation to institution to fund research staff (biostatistician).

NIH NEI K23EY033790 (M.Y.C.), Blind Children's Center (M.Y.C.), The Saban Research Institute (TSRI) at CHLA, Research to Prevent Blindness (M.Y.C., S.L.P.), NIH R21EY029655 (S.L.P.), NIH U10EY011751 (S.L.P.), NIH R01EY029687 (G.H.). NIH K08 EY030164 (E.D.G.). The funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the institutional review board at each separate institution. All research adhered to the tenets of the Declaration of Helsinki. Informed consent was waived by the institutional review board for this retrospective study.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Chang, Heidary, Beres, Pineles

Data collection: Chang, Heidary, Beres, Pineles, Gaier, Gise

Analysis and interpretation: Chang, Reid, Avramidis, Rostami, Narayanan

Obtained funding: N/A

Overall responsibility: Chang, Heidary, Beres, Pineles, Gaier, Gise, Reid, Avramidis, Rostami, Narayanan

Presented at the 2023 Annual meetings of the North American Neuro-Ophthalmology Society (NANOS, Orlando, FL, March 11-16, 2023) and American Association for Pediatric Ophthalmology and Strabismus (AAPOS, New York, NY, March 29-April 2, 2023).

Abbreviations and Acronyms:

AI = artificial intelligence; **AUC** = area under the receiver operating curve; **BONSAI** = Brain and Optic Nerve Study with Artificial Intelligence; **ICD** = International Classification of Diseases.

Keywords:
artificial intelligence, fundus photographs, papilledema, pediatric, pseudopapilledema.

Correspondence:
Melinda Y. Chang, MD, Department of Ophthalmology, Children's Hospital Los Angeles, Roski Eye Institute, Keck School of Medicine at the University of Southern California, 4650 Sunset Blvd. MS #88, Los Angeles, CA 90027. E-mail: melinda.y.wu@gmail.com.

References

1. Chang MY, Pineles SL. Optic disk drusen in children. *Surv Ophthalmol*. 2016;61:745−758.
2. Chang MY, Velez FG, Demer JL, et al. Accuracy of diagnostic imaging modalities for classifying pediatric eyes as papilledema versus pseudopapilledema. *Ophthalmology*. 2017;124:1839−1848.
3. Dahlmann-Noor AH, Adams GW, Daniel MC, et al. Detecting optic nerve head swelling on ultrasound and optical coherence tomography in children and young people: an observational study. *Br J Ophthalmol*. 2018;102:318−322.
4. Malem A, De Salvo G, West S. Use of MultiColor imaging in the assessment of suspected papilledema in 20 consecutive children. *J AAPOS*. 2016;20:532−536.
5. Martinez MR, Ophir A. Optical coherence tomography as an adjunctive tool for diagnosing papilledema in young patients. *J Pediatr Ophthalmol Strabismus*. 2011;48:174−181.
6. Thompson AC, Bhatti MT, El-Dairi MA. Bruch's membrane opening on optical coherence tomography in pediatric papilledema and pseudopapilledema. *J AAPOS*. 2018;22:38−43.e3.
7. Ozturk Z, Atalay T, Arhan E, et al. The efficacy of orbital ultrasonography and magnetic resonance imaging findings with direct measurement of intracranial pressure in distinguishing papilledema from pseudopapilledema. *Childs Nerv Syst*. 2017;33:1501−1507.
8. Chang MY, Binenbaum G, Heidary G, et al. Imaging methods for differentiating pediatric papilledema from pseudopapilledema: a report by the American Academy of Ophthalmology. *Ophthalmology*. 2020;127:1416−1423.
9. Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382:1687−1695.
10. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377−381.
11. Frisen L. Swelling of the optic nerve head: a staging scheme. *J Neurol Neurosurg Psychiatry*. 1982;45:13−18.
12. Avramidis K, Rostami M, Chang MY, Narayanan S. *Automating Detection of Papilledema in Pediatric Fundus Images with Explainable Machine Learning*. Bordeaux, France: IEEE International Conference on Image Processing (ICIP); 2022.
13. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Stat Med*. 2013;32:964−977.
14. Biousse V, Newman NJ, Najjar RP, et al. Optic disc classification by deep learning versus expert neuro-ophthalmologists. *Ann Neurol*. 2020;88:785−795.
15. Akbar S, Akram MU, Sharif M, et al. Decision support system for detection of papilledema through fundus retinal images. *J Med Syst*. 2017;41:66.
16. Fatima KN, Hassan T, Akram MU, et al. Fully automated diagnosis of papilledema through robust extraction of vascular patterns and ocular pathology from fundus photographs. *Biomed Opt Express*. 2017;8:1005−1024.
17. Ahn JM, Kim S, Ahn KS, et al. Accuracy of machine learning for differentiation between optic neuropathies and pseudopapilledema. *BMC Ophthalmol*. 2019;19:178.
18. Kovarik JJ, Doshi PN, Collinge JE, Plager DA. Outcome of pediatric patients referred for papilledema. *J AAPOS*. 2015;19:344−348.