Accuracy of a Deep Learning System for Classification of Papilledema Severity on Ocular Fundus Photographs

Caroline Vasseneix, MD, Raymond P. Najjar, PhD, Xinxing Xu, PhD, Zhiqun Tang, PhD, Jing Liang Loo, MBBS, MMed, FRCS(Ed), Shweta Singhal, MBBS, Sharon Tow, MBBS, Leonard Milea, Daniel Shu Wei Ting, MD, PhD, Yong Liu, PhD, Tien Y. Wong, MD, PhD, Nancy J. Newman, MD, Valerie Biousse, MD, and Dan Milea, PhD, on behalf of the BONSAI Group

Neurology® 2021;97:e369-e377. doi:10.1212/WNL.0000000000012226

Abstract

Objective
To evaluate the performance of a deep learning system (DLS) in classifying the severity of papilledema associated with increased intracranial pressure on standard retinal fundus photographs.

Methods
A DLS was trained to automatically classify papilledema severity in 965 patients (2,103 mydriatic fundus photographs), representing a multiethnic cohort of patients with confirmed elevated intracranial pressure. Training was performed on 1,052 photographs with mild/moderate papilledema (MP) and 1,051 photographs with severe papilledema (SP) classified by a panel of experts. The performance of the DLS and that of 3 independent neuro-ophthalmologists were tested in 111 patients (214 photographs, 92 with MP and 122 with SP) by calculating the area under the receiver operating characteristics curve (AUC), accuracy, sensitivity, and specificity. Kappa agreement scores between the DLS and each of the 3 graders and among the 3 graders were calculated.

Results
The DLS successfully discriminated between photographs of MP and SP, with an AUC of 0.93 (95% confidence interval [CI] 0.89–0.96) and an accuracy, sensitivity, and specificity of 87.9%, 91.8%, and 86.2%, respectively. This performance was comparable with that of the 3 neuro-ophthalmologists (84.1%, 91.8%, and 73.9%, p = 0.19, p = 1, p = 0.09, respectively). Misclassification by the DLS was mainly observed for moderate papilledema (Frisén grade 3). Agreement scores between the DLS and the neuro-ophthalmologists’ evaluation was 0.62 (95% CI 0.57–0.68), whereas the intergrader agreement among the 3 neuro-ophthalmologists was 0.54 (95% CI 0.47–0.62).

Conclusions
Our DLS accurately classified the severity of papilledema on an independent set of mydriatic fundus photographs, achieving a comparable performance with that of independent neuro-ophthalmologists.

Classification of Evidence
This study provides Class II evidence that a DLS using mydriatic retinal fundus photographs accurately classified the severity of papilledema associated in patients with a diagnosis of increased intracranial pressure.

From the Singapore Eye Research Institute (C.V., R.P.N., Z.T., J.L.L., S.S., S.T., D.S.W.T., T.Y.W., D.M.); Duke-NUS Medical School (R.P.N., J.L.L., S.S., S.T., T.Y.W., D.M.); Institute of High Performance Computing (X.X., Y.L.); Agency for Science, Technology and Research (A*STAR); Singapore National Eye Centre (J.L.L., S.S., S.T., D.S.W.T., T.Y.W., D.M.); University of Berkeley (L.M.), CA; Departments of Ophthalmology and Neurology (N.J.N., V.B.), Emory University School of Medicine, Atlanta, GA; and Copenhagen University Hospital (D.M.), Denmark.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

BONSAI Group coinvestigators are listed at http://links.lww.com/WNL/B432.

Copyright © 2021 American Academy of Neurology

Copyright © 2021 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.
Papilledema, defined as optic nerve head swelling associated with any cause of intracranial hypertension, can result in permanent vision loss.2,3 Papilledema severity at presentation is the most important prognostic factor for subsequent visual outcomes.3-9 Patients with severe papilledema may have progressive vision loss and visual field constriction due to retinal nerve fiber loss, thus requiring closer monitoring and more invasive treatment, whereas those with mild papilledema and no optic atrophy usually have good visual outcomes.5,8,9 However, the evaluation of papilledema severity, based on the 5-grade modified Frisén scale classification mainly used in clinical trials (with 1 being very mild papilledema and 5 very severe papilledema),10 is difficult to apply and subject to high variability.10-14 Hence, neurologists, especially those not confident in performing ophthalmoscopy,15 usually rely on ophthalmologists to determine the severity of papilledema.16,17 Fundus photography is now increasingly used in various clinical settings for screening purposes,18,19 and may be augmented with artificial intelligence deep learning techniques for automated image interpretation.20,21 Recently, the Brain and Optic Nerve Study with Artificial Intelligence (BONSAI) deep learning system was shown to accurately discriminate papilledema from normal and other abnormal optic discs on fundus photographs, with a performance comparable to that of expert neuro-ophthalmologists.22

The aim of the current study was to develop, train, and test a new DLS to automatically classify the severity of papilledema and to compare the performance of this DLS with the classification performance of 3 neuro-ophthalmologists.

**Methods**

This study, performed by the BONSAI Consortium,22 included investigators and patients from 14 countries. Our primary research questions, aiming to provide a Class II level of evidence, were the following: (1) Is a DLS capable of discriminating mild to moderate from severe papilledema on mydriatic fundus photographs? (2) Is the DLS’s performance in classifying the severity of papilledema on fundus photographs comparable to that of neuro-ophthalmologists?

**Standard Protocol Approvals, Registrations, and Patient Consents**

The study was approved by the Centralized Institutional Review Board of SingHealth, Singapore, and each contributing institution for any experiments using human subjects, and was conducted in accordance with the Declaration of Helsinki. Informed consent was exempted given the retrospective nature of the study and the use of de-identified ocular fundus photographs.

**Inclusion/Exclusion Criteria**

The study included unaltered, de-identified digital ocular fundus photographs obtained in patients with confirmed intracranial hypertension and papilledema from 19 neuro-ophthalmology centers participating in the BONSAI consortium.22 The fundus photographs, taken at various fields of view (20°–45°) including the optic disc, were obtained after pupillary dilation, using 15 different cameras, mydriatic or nonmydriatic, depending on the center (table 1).

Experts from each participating center provided photographs of patients with confirmed papilledema (criteria previously published23). Intracranial hypertension was confirmed in every patient by brain imaging (e.g., showing an intracranial mass or venous sinus thrombosis) or elevated CSF opening pressure and follow-up visits. Papilledema was diagnosed only if the optic disc swelling was related to confirmed raised intracranial pressure (ICP). Patients with a diagnosis of idiopathic intracranial hypertension met the modified Dandy criteria.24

The fundus photographs were divided into 2 datasets representing a mix of consecutive and convenience samples. The training dataset, used to train the DLS to classify the severity of papilledema, was composed of all papilledema images previously included in the training cohort of our first BONSAI study. The testing dataset was obtained by choosing randomly 222 images from 4 participating centers of the same study, independent from the training dataset, in order to test the performance of the DLS after training, and to test 3 independent neuro-ophthalmologists for comparison with the DLS.

Two experts (V.B., N.J.N.) independently reviewed all fundus photographs of the training (2,524 photographs) and testing (222 photographs) datasets, and classified papilledema severity according to a simple 2-grade classification (see below). Classification by the experts was performed under standard conditions on a computer screen (LG-34WK650, 100% brightness, 80% contrast) using semi-automated software developed by L.M. and R.P.N.25; the severity scores were automatically included into an Excel spreadsheet. In the case of discordance between the 2 experts, the classification was adjudicated by 2 additional neuro-ophthalmologists (D.M., C.V.), and a consensus was obtained for all images, used as reference standard. Only patients with active papilledema

---

**Glossary**

AUC = area under the receiver operating characteristic curve; BONSAI = Brain and Optic Nerve Study with Artificial Intelligence; CI = confidence interval; DLS = deep learning system; GCC = ganglion cell complex; ICP = intracranial pressure; IIH = idiopathic intracranial hypertension; OCT = optical coherence tomography.
were included, and optic discs with atrophic papilledema (defined as definite atrophy with no active swelling) were excluded from the study (394/2,524 [15.6%] photographs excluded in the training dataset and 8/222 [3.6%] in the testing dataset). Fundus photographs of insufficient quality were also excluded (27/2,524 [1.1%] in the training dataset, none in the testing dataset). A total of 2,103 and 214 fundus photographs were included in the training and testing datasets, respectively (figure 1).

### Papilledema Severity Classification
We created a simple 2-grade papilledema severity classification (figure 2): (1) mild to moderate papilledema, corresponding to Frisén grades 1–3, defined as disc edema with no obscuration of major blood vessels (arteries and veins) on the disc; (2) severe papilledema, that is, Frisén grades 4 and 5, defined as disc edema associated with any obscuration of major blood vessels (arteries and/or veins) on the disc. The presence of hemorrhages and exudates did not influence the classification, and optic discs with atrophic papilledema were excluded from the study.

### Study Population

#### Training Dataset
A total of 2,103 fundus photographs of 965 patients, 1,052 mild to moderate papilledema and 1,051 severe papilledema, were included in the training dataset, collected from 16 participating centers of BONSAI (table 2). A total of 685 patients had a photograph taken in both eyes and 146 in 1 eye.

#### Testing Dataset
The testing dataset included 214 photographs (92 MP, 122 SP) from 111 patients (103 patients with both eyes imaged and 8 with 1 eye), randomly collected from 4 participating centers (Bangkok, Thailand; Freiburg, Germany; Tehran, Iran; Angers, France) of the BONSAI study (table 2).

### Deep Learning System
Deep learning is a technique of machine learning, which consists of multiple layers of convolutional neural networks with the capability to learn image features and classify images without using hand-crafted features. DLs need to be trained on large datasets, and the classification performance is subsequently evaluated on part of the training dataset (validation) or on an independent external dataset (testing dataset).

In our study, the DLS was composed of a segmentation network and a classification network. First, the optic disc was automatically located on the fundus photograph by the segmentation network (U-Net) as previously described for the BONSAI-DLS. The segmentation network is based on U-
Net, which was widely used for biomedical image segmentation tasks. The U-Net was trained to localize the optic disc region based on a total of 6,370 fundus images with masks annotated in pixel level. The trained U-Net was then applied to test the full fundus image to generate the optic disc region automatically, which was further used as the input to the classification network.

Then, the classification network classified the optic disc into 1 of 2 classes: mild to moderate or severe papilledema (classification network, VGGNet). The classification network is based on the convolutional neural networks. At the last convolutional layer of the VGGNet, 2 dense layers were added with a SoftMax layer to obtain the 2-class outputs. The classification network was initialized using weights pretrained on ImageNet and fine-tuned in an end-to-end manner to achieve the optimal performance. The classification network was trained on 2,103 fundus photographs (training dataset) to automatically classify papilledema's severity into the 2 classes defined as the reference standard (mild to moderate and severe papilledema). The network weights were updated iteratively based on the difference signals between the outputs of DLS and the clinical reference standard on the severity levels using backpropagation algorithm. The trained segmentation

Figure 1 Flowchart of Inclusion and Exclusion of Fundus Photographs

![Flowchart of Inclusion and Exclusion of Fundus Photographs](image)

- Training dataset
  - Fundus photographs of papilledema from 16 participating centers (n = 2,524)
  - Excluded (n = 421):
    - Atrophic papilledema (394)
    - Insufficient quality (27)
  - Fundus photographs included (n = 2,103)

- Testing dataset
  - Fundus photographs of papilledema from 4 participating centers (n = 222)
  - Excluded (n = 8):
    - Atrophic papilledema (8)
  - Fundus photographs included (n = 214)

Process of inclusion and exclusion of fundus photographs in the training and testing datasets.

Figure 2 Papilledema Severity Classification

![Papilledema Severity Classification](image)

This figure represents the 2-grade papilledema severity classification system used in our study, separating mild to moderate papilledema (Frisén grade 1 to 3) and severe papilledema (Frisén grade 4 and 5). Mild to moderate papilledema was defined as disc edema with no obscuration of major blood vessels (arteries and veins) on the disc and severe papilledema as disc edema with any obscuration of major blood vessels on the disc.
network and the classification network were tested on another 214 images from the external testing dataset to get the prediction outcome. To report performance characteristics for this model, the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity were calculated. Diagnosis was provided at eye level for each included image.

Testing of 3 Independent Neuro-ophthalmologists for Comparison With the DLS
In order to compare the performance of the DLS with the classification performance of neuro-ophthalmologists, we tested 3 independent neuro-ophthalmologists (S.S., J.L.L., S.T.) who were asked to independently classify the papilledema severity into one of the 2 previously described classes, after a brief training session. The testing was performed on the same 214 images from the testing dataset used for the DLS.

For this purpose, images were presented to the 3 neuro-ophthalmologists on the same individual computer screen (LG-34WK650, 100% brightness, 80% contrast), using the semiautomated software described above. The neuro-ophthalmologists were masked to patients’ clinical information and to the classification assigned by the other evaluators and by the DLS. For comparisons and statistical analysis purposes, we used a majority agreement grade, defined as the severity classification reported by at least 2 of the 3 participating neuro-ophthalmologists, for each image.

Statistical Analysis
The performance characteristics of the DLS and of the 3 neuro-ophthalmologists were evaluated by calculating the AUC, sensitivity, specificity, and accuracy in the external testing dataset.

A pairwise McNemar test with Bonferroni correction was used to compare the accuracies, sensitivities, and specificities between the DLS and each neuro-ophthalmologist and between the DLS and the majority agreement among the 3 neuro-ophthalmologists.

Cohen kappa agreement scores were used for comparison between the DLS and each of the 3 neuro-ophthalmologists and between the DLS and the majority agreement among the 3 neuro-ophthalmologists. Kappa agreement scores were interpreted according to previously published scale (0–0.20: no agreement, 0.21–0.39: minimal; 0.40–0.59: weak; 0.60–0.79: moderate; 0.80–0.90: strong; >0.90: almost perfect). A p value less than 0.05 was considered statistically significant.

Data Availability
Anonymized data will be shared by justified request from any qualified investigator.

Results
Patient and Image Characteristics
The final training dataset included 965 patients (representing 1777 eyes and 2,103 images); 397 patients (731 eyes) had mild to moderate papilledema, 431 (772 eyes) had severe papilledema, and 137 (274 eyes) had mild to moderate papilledema in one eye and severe papilledema in the other eye. Among those 965 patients, 822 (85%) had papilledema due to idiopathic intracranial hypertension (IIH) and 143 (15%) presented with secondary causes of intracranial hypertension such as cerebral venous sinus thrombosis, meningitis, or brain tumor. Patient demographics and image characteristics are described in table 3.

The testing dataset included 111 patients (representing 214 eyes or images), of whom 40 (77 eyes or images) had mild to moderate papilledema, 56 (107 eyes or images) had severe papilledema, and 15 (20 eyes or images) had moderate to severe papilledema. The patient demographics are described in table 3.

Table 2: List of Participating Centers and Number of Photographs Included per Center

<table>
<thead>
<tr>
<th>Participating center</th>
<th>Number of photographs included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary training dataset</td>
<td></td>
</tr>
<tr>
<td>Angers, France</td>
<td>286</td>
</tr>
<tr>
<td>Atlanta, USA</td>
<td>1,157</td>
</tr>
<tr>
<td>Baltimore, USA</td>
<td>136</td>
</tr>
<tr>
<td>Bologna, Italy</td>
<td>11</td>
</tr>
<tr>
<td>Bordeaux, France</td>
<td>26</td>
</tr>
<tr>
<td>Chennai, India</td>
<td>183</td>
</tr>
<tr>
<td>Coimbra, Portugal</td>
<td>17</td>
</tr>
<tr>
<td>Freiburg, Germany</td>
<td>10</td>
</tr>
<tr>
<td>Geneva, Switzerland</td>
<td>12</td>
</tr>
<tr>
<td>Grenoble, France</td>
<td>11</td>
</tr>
<tr>
<td>Hong Kong, China</td>
<td>4</td>
</tr>
<tr>
<td>London, UK</td>
<td>35</td>
</tr>
<tr>
<td>Manila, Philippines</td>
<td>18</td>
</tr>
<tr>
<td>Paris, France</td>
<td>86</td>
</tr>
<tr>
<td>Singapore</td>
<td>18</td>
</tr>
<tr>
<td>Sydney, Australia</td>
<td>65</td>
</tr>
<tr>
<td>Syracuse, USA</td>
<td>28</td>
</tr>
<tr>
<td>Total training dataset</td>
<td>2,103</td>
</tr>
<tr>
<td>External testing dataset</td>
<td></td>
</tr>
<tr>
<td>Angers, France</td>
<td>10</td>
</tr>
<tr>
<td>Bangkok, Thailand</td>
<td>86</td>
</tr>
<tr>
<td>Freiburg, Germany</td>
<td>35</td>
</tr>
<tr>
<td>Tehran, Iran</td>
<td>83</td>
</tr>
<tr>
<td>Total testing dataset</td>
<td>214</td>
</tr>
</tbody>
</table>
papilledema, and 15 (30 eyes or images) had mild to moderate papilledema in one eye and severe papilledema in the other eye. Seventy-five patients (68%) had papilledema from IIH and the remaining from secondary causes.

**Papilledema Severity Classification by the DLS and Neuro-ophthalmologists**

In the testing dataset, the DLS successfully discriminated severe from mild to moderate papilledema with an AUC of 0.93 (95% confidence interval [CI] 0.89–0.96), an accuracy of 87.9% (95% CI 82.7%–91.9%), a sensitivity of 91.8% (95% CI 86.9%–96.7%), and a specificity of 82.6% (95% CI 74.9%–90.4%) (figures 3 and 4). Among the 26 images misclassified by the DLS, 16 cases of mild to moderate papilledema were misclassified as severe; 14 of these 16 (87.5%) were photographs of papilledema Frisén grade 3, and 2 of them Frisén grade 2; 10 images of severe papilledema were misclassified as mild to moderate, all Frisén grade 4 with minimal vessel obscuration on the disc or with hemorrhages and cotton-wool spots on the disc (eFigures 1 and 2, data available from Dryad, doi.org/10.5061/dryad.66t1g1k1x). The ability of the 3 independent neuro-ophthalmologists to discriminate severe from mild to moderate papilledema was comparable to the DLS, with an accuracy of the majority agreement among neuro-ophthalmologists of 84.1% (95% CI 78.5%–88.7%, $p = 0.19$), a sensitivity of 91.8% (95% CI 86.9%–96.7%, $p = 1$), and a specificity of 73.9% (95% CI 64.9%–82.9%, $p = 0.09$) (figures 3 and 4).

Agreement scores between the DLS and neuro-ophthalmologist 1, 2, and 3 were 0.72 (95% CI 0.67–0.76), 0.43 (95% CI 0.37–0.49), and 0.60 (95% CI 0.55–0.65), respectively, and between the DLS and the majority agreement of neuro-ophthalmologists, 0.62 (95% CI 0.57–0.68). A weak intergrader agreement of 0.54 (95% CI 0.47–0.62) was found among the 3 neuro-ophthalmologists. Disagreement among neuro-ophthalmologists was observed for 67 photographs, among them 46 (68.6%) photographs of moderate papilledema with a Frisén grade of 3, the other 21 photographs of severe papilledema with a Frisén grade of 4.

**Table 3** Demographics and Image Characteristics in the Training and Testing Datasets

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Training dataset</th>
<th>Testing dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients (images)</td>
<td>965 (2,103)</td>
<td>111 (214)</td>
</tr>
<tr>
<td>Age, y, mean (SD)</td>
<td>31.7 (12.8)</td>
<td>32.0 (15.4)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>715 (74.9)</td>
<td>77 (69.4)</td>
</tr>
<tr>
<td>Images with mild to moderate papilledema, n (%)</td>
<td>1,052 (50.0)</td>
<td>92 (43.0)</td>
</tr>
<tr>
<td>Images with severe papilledema, n (%)</td>
<td>1,051 (50.0)</td>
<td>122 (57.0)</td>
</tr>
</tbody>
</table>

**Discussion**

In this study, a DLS trained on 2,103 ocular fundus photographs to classify the severity of papilledema from intracranial hypertension discriminated between mild to moderate and severe papilledema on an independent testing dataset of 214 fundus photographs, with a comparable performance to that of 3 neuro-ophthalmologists.

Papilledema\textsuperscript{33,34} is the only objective clinical sign of intracranial hypertension.\textsuperscript{35,36} Because the degree of papilledema at presentation is a reliable indicator of subsequent visual outcomes from secondary optic atrophy,\textsuperscript{3–9} the severity of papilledema influences the management and the frequency of visual monitoring during follow-up of patients with elevated ICP.\textsuperscript{37,38} Our simple binary classification of papilledema severity aimed at identification of patients with lower risk (mild to moderate papilledema) or higher risk (severe papilledema) would be more useful for clinical decision-making.
papilledema) of visual loss. Hence, a patient with mild to moderate papilledema from IIH might be managed with weight loss and oral medications as an outpatient, whereas a patient with severe papilledema should have closer follow-up, sometimes inpatient, and benefit from timely surgical intervention, especially in those cases with visual loss. A DLS, capable of automatically classifying the severity of papilledema on fundus photographs, could assist nonexperts with a more accurate prognosis, treatment strategy, and effect of treatment, as a diagnostic or prognostic tool along with the clinical examination and additional tests such as computer-assisted perimetry or optical coherence tomography (OCT).

It could be argued that ophthalmology consultation would obviate the need for automated fundus photographic interpretation by a DLS. However, despite the simplification of the papilledema severity classification in our study, the intergrader agreement among the 3 neuro-ophthalmologists was relatively weak (0.54), confirming the variability of human evaluation of papilledema severity, particularly for moderate papilledema (Frisén grade 3). Nevertheless, our DLS could accurately distinguish between mild to moderate and severe papilledema on fundus photographs. The performance of the DLS was similar to that of 3 independent neuro-ophthalmologists with a comparable accuracy and sensitivity, and with a nonsignificantly higher specificity (82.6% for the DLS vs 73.9% for the majority agreement among the 3 neuro-ophthalmologists, \( p = 0.09 \)). The agreement scores between the DLS and the majority agreement among the 3 neuro-ophthalmologists (\( \kappa = 0.62 \)) was higher than the intergrader agreement among the 3 neuro-ophthalmologists (\( \kappa = 0.54 \)). Moderate intergrader agreement scores were also observed in studies involving glaucoma or retinal diseases, for example, for glaucomatous damage assessment of the optic disc under stereoscopic conditions by 6 glaucoma experts (\( \kappa = 0.50 \)), or for plus disease retinopathy of prematurity diagnosis by 9 experts (\( \kappa = 0.59 \) to 0.92). Similar results were previously described with a machine learning technique in a small study, which showed that a computer-aided image analyses, used to analyze features of papilledema on fundus photographs, could automatically grade papilledema with a substantial agreement with one expert grading (\( \kappa = 0.71 \)).

Our results might have implications for the management of raised ICP and papilledema in the future. However, further prospective validation studies are needed, at best in non-ophthalmologic clinical settings (i.e., neurology clinics, neurosurgery clinics, or emergency departments), and ideally using nonmydriatic digital cameras. If those studies confirm its applicability, a DLS, connected to a camera on site or remotely, could be used for the assessment of papilledema severity.

Our study has inherent limitations. As it was a retrospective data collection, visual function was not available and objective data such as OCT retinal nerve fiber thickness or macular ganglion cell complex (GCC) analysis were not systematically collected. Moreover, atrophic papilledema was excluded from this study, as we only trained the DLS to grade the severity of papilledema, not to identify associated atrophy, a difficult assessment even for the most experienced ophthalmologists. Hence, some of the patients with longstanding raised ICP who had both atrophy and residual papilledema at first presentation were excluded. In a future project, GCC-OCT,

---

**Figure 4** Performance (Accuracy, Sensitivity, Specificity) of the Deep Learning System (DLS) and the 3 Neuro-ophthalmologists in Classifying Papilledema Severity

Comparison of the performance (accuracy, sensitivity, specificity) of the DLS to (A) the detailed performance of each grader and (B) the performance of the majority agreement among the 3 graders in discriminating severe papilledema (SP) from mild to moderate papilledema (MP) on 214 fundus photographs (92 MP and 122 SP). Statistical significance of DLS vs human graders \(*p < 0.05\); **p < 0.01; ***p < 0.001. Data are represented as mean (95% confidence interval [CI]). 95% CIs were calculated using the asymptotic method.
which has proven useful in the detection of optic atrophy associated with papilledema,\textsuperscript{43,44} could also be incorporated into the DLS strategy, as already done in some glaucoma studies.\textsuperscript{45} We used a simplified classification of papilledema into 2 grades, instead of the 5 grades used in the Frisén scale. The Frisén scale is notoriously difficult to use, especially when attempting to differentiate grades 3 and 4.\textsuperscript{13,14} The few cases of misclassification by the DLS were mainly observed for moderate papilledema (Frisén grade 3) or for Frisén grade 4 papilledema with mild vessel obscuration on the optic disc or associated with hemorrhages or cotton-wool spots. Those images were challenging to classify by the experts as well. The lack of reproducibility and the inability of the Frisén scale to discern optic disc changes over time was demonstrated by Sinclair et al.,\textsuperscript{13} who proposed an alternative ranking of optic disc appearance related to papilledema severity that incorporates the development of secondary optic atrophy, with an improvement of complete agreement among reviewers from 1.6% of photographs with Frisén scale optic atrophy, with an improvement of complete agreement from 55% to 73%.\textsuperscript{14} Our simplified binary classification was designed to signal the presence of severe papilledema, a finding that should influence the acute management of these patients by nonophthalmologists.

We developed, trained, and tested a DLS that accurately discriminated mild to moderate from severe papilledema on mydriatic fundus photographs. In a subsequent comparison, the DLS had a performance comparable to 3 independent neuro-ophthalmologists. The automated recognition of severe papilledema by a DLS could be helpful in neurology, neurosurgery, and emergency settings for the management of patients with raised ICP. Additional prospective studies are needed to confirm the applicability of this DLS in real-life clinical settings.

**Study Funding**

Singapore National Medical Research Council (Clinician Scientist Individual Research grant CIRG18Nov-0013), The Duke-NUS Medical School, Ophthalmology and Visual Sciences Academic Clinical Program grant (05/FY2019/P2/06-A60) departmental grant, NIH/NEI core grant P30-EY06360 (Department of Ophthalmology, Emory University School of Medicine, Atlanta, GA), NIH/NINDS (RO1NSO89694).

**Disclosure**


**Publication History**

Received by Neurology November 9, 2020. Accepted in final form April 19, 2021.

---

**Appendix 1 Authors**

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caroline Vasseneix, MD</td>
<td>Singapore Eye Research Institute</td>
<td>Designed and conceptualized study, major role in the acquisition of data, analyzed and interpreted the data, drafted the manuscript for intellectual content</td>
</tr>
<tr>
<td>Raymond P. Najjar, PhD</td>
<td>Singapore Eye Research Institute; Duke-NUS Medical School, Singapore</td>
<td>Designed and conceptualized study, analyzed and interpreted the data, revised the manuscript for intellectual content</td>
</tr>
<tr>
<td>Xinxing Xu, PhD</td>
<td>Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore</td>
<td>Major role in the acquisition of data, interpreted the data, revised the manuscript for intellectual content</td>
</tr>
<tr>
<td>Zhiqun Tan, PhD</td>
<td>Singapore Eye Research Institute</td>
<td>Interpreted the data</td>
</tr>
<tr>
<td>Jing Liang Loo, MBBS, MMed, FRCS(Ed)</td>
<td>Singapore National Eye Centre</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Shweta Singhal, MBBS, PhD</td>
<td>Singapore National Eye Centre; Singapore Eye Research Institute</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Sharon Tow, MBBS, FRCS(Ed)</td>
<td>Singapore National Eye Centre; Duke-NUS Medical School</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Leonard Milea</td>
<td>UC Berkeley, CA</td>
<td>Design and conceptualized study, major role in the acquisition of data</td>
</tr>
<tr>
<td>Daniel Shu Wei Ting, MD, PhD</td>
<td>Singapore National Eye Centre; Singapore Eye Research Institute</td>
<td>Design and conceptualized study</td>
</tr>
<tr>
<td>Yong Liu, PhD</td>
<td>Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore</td>
<td>Major role in the acquisition of data, interpreted the data, revised the manuscript for intellectual content</td>
</tr>
<tr>
<td>Tien Y. Wong, MD, PhD</td>
<td>Singapore National Eye Centre; Singapore Eye Research Institute; Duke-NUS Medical School</td>
<td>Revised the manuscript for intellectual content</td>
</tr>
<tr>
<td>Nancy J. Newman, MD</td>
<td>Ophthalmology and Neurology Department, Emory University School of Medicine, Atlanta, GA</td>
<td>Designed and conceptualized study, major role in the acquisition of data, interpreted the data, revised the manuscript for intellectual content</td>
</tr>
<tr>
<td>Valerie Bioussé, MD</td>
<td>Ophthalmology and Neurology Department, Emory University School of Medicine, Atlanta, GA</td>
<td>Designed and conceptualized study, major role in the acquisition of data, interpreted the data, revised the manuscript for intellectual content</td>
</tr>
</tbody>
</table>
### Appendix 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dan Milea, MD, PhD</td>
<td>Singapore National Eye Centre; Singapore Eye Research Institute; Duke-NUS Medical School, Singapore; Copenhagen University Hospital, Denmark</td>
<td>Designed and conceptualized study, major role in the acquisition of data, interpreted the data, revised the manuscript for intellectual content</td>
</tr>
</tbody>
</table>

### Appendix 2

Coinvestigators are listed at links.lww.com/WNL/8432.

### References


Copyright © 2021 American Academy of Neurology. Unauthorized reproduction of this article is prohibited.
Accuracy of a Deep Learning System for Classification of Papilledema Severity on Ocular Fundus Photographs

Caroline Vasseneix, Raymond P. Najjar, Xinxing Xu, et al.

Neurology 2021;97:e369-e377 Published Online before print May 19, 2021
DOI 10.1212/WNL.0000000000012226

This information is current as of May 19, 2021