



UNIVERSITÉ  
DE GENÈVE

Archive ouverte UNIGE

<https://archive-ouverte.unige.ch>

Article scientifique

Article

2020

Appendix

Open Access

This file is a(n) Appendix of:

Artificial intelligence to detect papilledema from ocular fundus photographs

Milea, Dan; Biousse, Valérie; Bonzai, group

Collaborators: Sanda, Nicolae; Thumann, Gabriele

This publication URL:

<https://archive-ouverte.unige.ch/unige:155363>

Publication DOI:

[10.1056/NEJMoa1917130](https://doi.org/10.1056/NEJMoa1917130)

© This document is protected by copyright. Please refer to copyright holders for terms of use.

## Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Milea D, Najjar RP, Jiang Z, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. N Engl J Med 2020;382:1687-95. DOI: 10.1056/NEJMoa1917130

(PDF updated May 6, 2020)

## Supplementary Appendix

### Artificial Intelligence to Detect Papilledema From Ocular Fundus Photographs (BONSAI - Brain and Optic Nerve Study with Artificial Intelligence)

#### Table of Contents:

|  |           |
|--|-----------|
| <b>Section S1: List of Investigators.....</b>  | <b>3</b>  |
| BONSAI Study Group.....  | 3         |
| List of all participating centers (alphabetical order by city) .....   | 4         |
| <b>Section S2: Additional Methods.....</b>   | <b>9</b>  |
| a) Selection of participants and participants' characteristics.....  | 9         |
| b) List of digital retinal cameras used in each participating center.....  | 13        |
| c) Model development: Technical aspects.....   | 13        |
| d) Statistical and bootstrapping procedures.....   | 19        |
| <b>Section S3: Additional Analyses.....</b>  | <b>20</b> |
| a) Errors in classification by the deep learning system.....   | 20        |
| b) Labelling errors.....   | 20        |
| c) Prevalence of abnormal conditions and predictive values of the deep learning system.....  | 20        |
| <b>Section S4: Supplementary Figures S1-S6.....</b>  | <b>22</b> |
| Figure S1: Flow chart showing the process for inclusion and exclusion of ocular fundus<br>photographs.....   | 22        |
| Figure S2: Examples of fundus photographs and corresponding heatmaps.....  | 23        |
| Figure S3: Receiver operating characteristic curves and areas under the curves (AUC) of<br>individual folds for the 5-fold cross-validation performed on the primary dataset.....            | 26        |
| Figure S4: Technical model with segmentation and classification networks.....  | 26        |
| Figure S5: Predictive values of the deep learning system across a full prevalence range for the<br>detection of normal disks, disks with papilledema and other optic disk abnormalities..... | 27        |
| Figure S6: Errors in classification by the deep learning system.....   | 28        |
| <b>Section S5: Supplementary Tables S1-S6.....</b>   | <b>30</b> |

|   |           |
|---|-----------|
| Table S1: List of digital retinal cameras used in each participating center.....  | 30        |
| Table S2: Demographic distribution of patients from the training and external testing datasets.....                               | 31        |
| Table S3: Classification performance of single neural networks against a combination of two networks.....                         | 32        |
| Table S4: Classification performance of the deep learning system on the individual external testing datasets .....                | 33        |
| Table S5: Multi-class AUC for the 5 external testing datasets and comparison with the range of AUCs for one-vs-rest strategy..... | 35        |
| Table S6: Estimated prevalence of abnormal conditions from testing sites.....   | 36        |
| <b>Section S6: References for Supplementary Appendix.....</b>   | <b>37</b> |

## SECTION S1: LIST OF INVESTIGATORS

### BONSAI STUDY GROUP

This international consortium (BONSAI – Brain and Optic Nerve Study with Artificial Intelligence) was specifically created for the purpose of this study, with the contribution of numerous recognized neuro-ophthalmologists considered as experts world-wide. This consortium was formed by an initial steering committee, consisting of two groups of experts: 1) clinical neuro-ophthalmologists and neuroscientists; and 2) artificial intelligence experts.

#### **a) The initial steering committee members are listed below:**

##### 1. Neuro-ophthalmology

- a. Dan Milea MD, PhD (DM), Neuro-Ophthalmology department (Singapore National Eye Centre) and Head of the Visual Neuroscience Group, Singapore Eye Research Institute, Duke-NUS Medical School, Singapore
  - i. Principal Investigator and overall coordinator of the consortium
- b. Valérie Biousse, MD (VB) and Nancy J. Newman, MD (NJN), Neuro-Ophthalmologists (Atlanta, Georgia, USA)
  - i. Joint neuro-ophthalmology clinical leads of the consortium
- c. Raymond P. Najjar, PhD, Neuroscientist, Singapore Eye Research Institute and Duke-NUS Medical School, Singapore
  - i. Scientific and methodological lead of the study
- d. DM, VB and NJN are world experts in neuro-ophthalmology, with a special interest in optic nerve head abnormalities and ocular fundus imaging.

##### 2. Artificial intelligence experts

- a. Tien Y. Wong, MD, PhD (TYW), Daniel Ting, MD, PhD (DT) and Yong Liu, PhD (YL)
  - i. Joint technical leads
  - ii. Jointly, they have published >30 artificial intelligence ophthalmology articles in JAMA, Nature Medicine, Nature Biomedical Engineering, Lancet Digital Health, Nature Digital Medicine, Progress in Retinal and Eye Research, Ophthalmology, JAMA Ophthalmology, etc.
  - iii. DT - Founding executive committee member, American Academy of Ophthalmology Artificial Intelligence Taskforce
  - iv. Editors for artificial intelligence and data science in 4 high impact journals (TYW - Lancet Digital Health, JAMA Ophthalmology; DT - Ophthalmology and British Journal of Ophthalmology)

The 24 local site principal investigators (PIs) were selected based on their recognized expertise in ophthalmology and neuro-ophthalmology and on their ability to contribute to the consortium's collection with reliable and adequate numbers of fundus photographs (and the corresponding clinical information and reference standards). Each site included at least one internationally recognized senior neuro-ophthalmologist able to provide ocular fundus photographs associated with a definite clinical diagnosis (reference standard).

**b) BONSAI Study Group: List of all participating centers (alphabetical order by city):**

**ANGERS, France:**

Department of Ophthalmology

University Hospital Angers, Angers, France

1/ **Philippe Gohier, MD**-phgohier@chu-angers.fr

2/ **Barnabé Rondet-Courbis, MD** -barnabe.rondet@orange.fr

**ATLANTA, GA, USA**

Departments of Ophthalmology, Neurology and Neurological Surgery

Emory University School of Medicine, Atlanta, GA, USA

1/ **Valérie Biousse, MD** -vbiousse@emory.edu

2/ **Nancy J. Newman, MD**-ophntn@emory.edu

3/ **Caroline Vasseneix, MD**-caroline.vasseneix@emory.edu

**BALTIMORE, MD, USA**

Departments of Ophthalmology, Neurology and Neurosurgery

Johns Hopkins University School of Medicine, Baltimore, MD, USA

1/ **Neil Miller, MD**-nrmiller@jhmi.edu

**BANGKOK, Thailand:**

Department of Ophthalmology

Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

1/ **Tanyatuth Padungkiatsagul, MD**-Blu\_c16@hotmail.com

2/ **Anuchit Poonyathalang, MD**-Au.tumn@yahoo.com

3/ **Yanin Suwan, MD**-Yanin.suwan@gmail.com

4/ **Kavin Vanikieti, MD**-Vanikieti.kavin@gmail.com

**BOLOGNA, Italy:**

1/ **Giulia Amore, MD**-amoregiulia@hotmail.it

Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna

**2/ Piero Barboni, MD**-p.barboni@studiodazeglio.it

Studio Oculistico D'Azeglio, Bologna, Italy

**3/ Michele Carbonelli, MD**-m.carbonelli@studiodazeglio.it

Studio Oculistico D'Azeglio, Bologna, Italy

**4/ Valerio Carelli, MD, PhD**-valerio.carelli@unibo.it

IRCCS Istituto delle Scienze Neurologiche di Bologna, UOC Clinica Neurologica, Bologna, Italy

Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, Bologna, Italy

**5/ Chiara La Morgia, MD, PhD**-chiara.lamorgia@unibo.it

IRCCS Istituto delle Scienze Neurologiche di Bologna, UOC Clinica Neurologica, Bologna, Italy

Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, Bologna, Italy

**6/ Martina Romagnoli, PhD**-martina.romagnoli87@gmail.com

IRCCS Istituto delle Scienze Neurologiche di Bologna, UOC Clinica Neurologica, Bologna, Italy

#### **BORDEAUX, France:**

Service d'Ophtalmologie. Unité Rétine - Uvéites - Neuro-Ophtalmologie

Hôpital Pellegrin, CHU de Bordeaux, Bordeaux, France

**1/ Marie-Bénédicte Rougier, MD, PhD**-marie-benedicte.rougier@chu-bordeaux.fr

#### **CHENNAI, India:**

Dept of Neuro-ophthalmology. Sankara Nethralaya-A unit of Medical Research Foundation, Chennai, India

**1/ Selvakumar Ambika, DO,DNB**-drsa@snmail.org

**2/ Swetha Komma, DO,DNB**-kommaswetha2@gmail.com

#### **COIMBRA, Portugal:**

Department of Ophthalmology, Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra, Portugal

Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), Faculty of Medicine University of Coimbra (FMUC), Coimbra, Portugal

**1/ Pedro Fonseca, MD**-pedroluisfonseca@gmail.com

**2/ Miguel Raimundo, MD**-mglraimundo@gmail.com

#### **COPENHAGEN, Denmark:**

Department of Ophthalmology, Rigshospitalet, University of Copenhagen, Glostrup, Denmark

**1/ Steffen Hamann, MD, PhD**-steffen.hamann@regionh.dk

**2/ Isabelle Karlesand, MD**-anna.isabelle.wanda.karlesand@regionh.dk

#### **FREIBURG, Germany:**

Eye Center, Medical Center, Medical Faculty, University of Freiburg, Freiburg, Germany (W.A.L.)

**1/ Lars Fuhrmann** -fuhr.lars@gmail.com

**2/ Sebastian Kuechlin, MD**-sebastian.kuechlin@uniklinik-freiburg.de

**3/ Wolf Alexander Lagreze, MD**-wolf.lagreze@uniklinik-freiburg.de

**GENEVA, Switzerland:**

The Department of Clinical Neuroscience, Geneva University Hospital, Geneva, Switzerland

**1/ Nicolae Sanda, MD, PhD**-nicolae.sanda@hcuge.ch

**2/ Gabriele Thumann, MD, PD**-gabriele.thumann@hcuge.ch

**GRENOBLE, France:**

Department of Ophthalmology, University Hospital of Grenoble-Alpes,  
and Grenoble-Alpes University, HP2 Laboratory, INSERM U1042, Grenoble, France

**1/ Florent Aptel, MD, PhD**-aptel\_florent@hotmail.com

**2/ Christophe Chiquet, MD, PhD**-christophe.chiquet@inserm.fr

**GUANGZHOU, China:**

Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, P.R. China

**1/ Kaiqun Liu, MD**-liukaiqun\_cm@163.com

**2/ Hui Yang MD, PhD**-13710584767@163.com

**HONG KONG, China:**

Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China

Hong Kong Eye Hospital, Hong Kong, China

**1/ Carmen KM Chan, MRCP, FRCSEd(Ophth)**-ckmc01@ha.org.hk

**2/ Noel CY Chan, FRCSEd(Ophth)**-ccy178@ha.org.hk

**3/ Carol Y Cheung, PhD**-carolcheung@cuhk.edu.hk

**LILLE, France:**

Department of Ophthalmology, Lille Catholic Hospital, Lille Catholic University and Inserm U1171, Lille, France

**Thi Ha Chau Tran, MD**-tran.hachau@ghicl.net

**LONDON, United Kingdom:**

**1/ James Acheson, BM, MRCP (UK), FRCOphth**-james.acheson1@nhs.net

Moorfields Eye Hospital NHS Foundation Trust, London, UK

National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS trust, London, UK.

**2/ Maged S Habib, MSc Ophth, FRCS, FRCOphth, MD**-maged.habib@chsft.nhs.uk

South Tyneside and Sunderland NHS Foundation Trust. Sunderland, UK



**3/ Neringa Jurkute, MD, FEBO-n.jurkute@nhs.net**

Moorfields Eye Hospital NHS Foundation Trust, London, UK

UCL Institute of Ophthalmology, University College London, London, UK

**4/ Patrick Yu-Wai-Man, MB, BS, FRCPath, FRCOphth, PhD-p.yu-wai-man@nhs.net**

Moorfields Eye Hospital NHS Foundation Trust, London, UK

UCL Institute of Ophthalmology, University College London, London, UK

Cambridge Eye Unit, Addenbrooke's Hospital, Cambridge University Hospitals, Cambridge, UK

Cambridge Centre for Brain Repair and MRC Mitochondrial Biology Unit, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

**MANILA, Philippines:**

American Eye Center, Mandaluyong City, Philippines

**Richard Kho, MD-rich\_kho@yahoo.com**

**MANNHEIM, Germany:**

Department of Ophthalmology, Medical Faculty Mannheim of the Ruprecht-Karls-University of Heidelberg, Mannheim, Germany

**Jost B Jonas, MD-jost.jonas@medma.uni-heidelberg.de**

**MAYO-CLINIC, ROCHESTER, MN, USA:**

Department of Ophthalmology and Neurology, Mayo Clinic, Rochester, MN, USA

**1/ John J. Chen, MD, PhD-chen.john@mayo.edu**

**2/ Nouran Sabbagh, MD-Sabbagh.Nouran@mayo.edu**

**PARIS, ROTHSCHILD FOUNDATION HOSPITAL, France:**

Fondation Adolphe de Rothschild, Paris, France

**1/ Catherine Clermont-Vignal, MD-cvignal@for.paris**

**2/ Rabih Hage, MD-rhage@for.paris**

**3/ Raoul Kanav Khanna, MD-krkhanna1@gmail.com**

**SNEC, SINGAPORE:**

Singapore National Eye Centre

Singapore Eye Research Institute, Singapore

Duke-NUS Medical School, Singapore

Yong Loo Lin School of Medicine, National University of Singapore

**1/ Tin Aung, MD, PhD-aung.tin@singhealth.com.sg**

**2/ Ching-Yu Cheng, MD, PhD-cheng.ching.yu@seri.com.sg**

**3/ Ecosse Lamoureux, MSc, PhD-ecosse.lamoureux@seri.com.sg**

**4/ Jing Liang Loo, MBBS, MMed, FRCS(Ed)-loo.jing.liang@singhealth.com.sg**

**5/ Dan Milea, MD, PhD-dan.milea@sneec.com.sg**

- 6/ Raymond P. Najjar, PhD**-raymond.najjar@seri.com.sg
- 7/ Leopold Schmetterer, PhD**-leopold.schmetterer@seri.com.sg
- 8/ Shweta Singhal, MBBS, PhD**-Shweta.singhal@snec.com.sg
- 9/ Daniel Ting, MD, PhD**-daniel.ting.s.w@singhealth.com.sg
- 10/ Sharon Tow, MBBS, FRCSEd**-sharon.tow.l.c@snec.com.sg
- 11/ Caroline Vasseneix, MD**-caroline.vasseneix@seri.com.sg
- 12/ Tien Yin Wong, MD, PhD**-wong.tien.yin@singhealth.com.sg

**SINGAPORE, IHPC, AStar:**

Institute of High Performance Computing, Agency for Science, Technology and Research (A\*STAR),  
Singapore

- Yong Liu, PhD**-liuyong@ihpc.a-star.edu.sg
- Xinxing Xu, PhD**-xuxinx@ihpc.a-star.edu.sg
- Zhubo Jiang, MSc**-jiangzhubo1992@outlook.com

**SYDNEY, Australia:**

Save Sight Institute, Faculty of Health and Medicine, The University of Sydney, NSW Australia.

- 1/ Clare L Fraser, MBBS, MMed, FRANZCO**-clare.fraser@sydney.edu.au

**SYRACUSE, NY, USA:**

Department of Neurology, SUNY Upstate Medical University, Syracuse, NY 13210

- 1/ Luis J. Mejico, MD**-mejicol@upstate.edu
- 2/ Andrew L. Orenberg, MD**-orenbera@upstate.edu

**TEHERAN, Iran:**

Farabi Eye Hospital, Tehran University of Medical Science, Tehran, Iran, 13366-16351

- 1/ Masoud Aghsaei Fard, MD**-masood219@gmail.com

## SECTION S2: ADDITIONAL METHODS

### **a) Selection of Participants and Participants' Characteristics**

#### **a-1. Participating Centers and Datasets:**

All participants included in this study were either recruited from reference neuro-ophthalmology clinics or population-based studies. We selected participating centers from a large number of countries in order to obtain a representative set of fundus photographs of patients with a variety of optic nerve head disorders, occurring in a wide range of patient ethnicities and ages. This was essential in order to show the reproducibility of the deep learning system in geographically distinct populations. Neuro-ophthalmologists routinely obtain fundus photographs on patients with abnormal optic nerves and often on normal patients, similar to primary care physicians measuring blood pressure on nearly all patients, for example. Not all neuro-ophthalmologists do that and this is why we only included centers that routinely obtain fundus photographs on most patients (see list of centers in Section S1 above).

The BONSAI consortium collected raw data over 12 months. We utilized the first 19 datasets gathered over the initial 9 months for training and validation of the deep learning system. Subsequently, the trained model was tested on 5 independent external testing datasets, provided by 5 additional neuro-ophthalmology centers with well-established clinical and imaging protocols for patients with optic disk abnormalities. These 5 external testing centers were chosen because of their comparable sizes (200-300 images), well-distributed diagnoses across each center, and the fact that they represented various ethnicities from different continents to confirm applicability and generalization of the deep learning system.

As expected with a retrospective collection of fundus photographs from multiple centers, there is considerable variability in the proportion of abnormals and normals at each site. This is a result of the specific interests and expertise of each participating center. This should have no impact on the results of our study (see details in Section S2c below; technical aspects, handling class imbalance).

#### **a-2. Patient Selection and Inclusion Criteria (participating centers):**

The main inclusion criteria were fundus photographs of the optic disks and definite corresponding diagnoses clinically made by expert neuro-ophthalmologists in reference centers.

Some of the centers provided consecutive series of patients/photographs either obtained routinely in clinic (such as Emory, USA or Copenhagen, Denmark) or from previously generated large datasets of photographs (such as India and Singapore). Others provided previously collected samples

of photographs for each of the optic nerve disorders based on their local practice and expertise in order to provide a variety of different optic nerve pathologies and normal optic nerves, and therefore could be considered “convenience samples”. However, such mix of samples should not be an issue for this specific study which is not about the relative percentages of normals and pathologies in the general population, but rather about teaching the deep learning system machine to recognize normals and various optic nerve pathologies. It was essential to gather a large number of well-defined high-quality photographs of various optic nerve appearances correctly labelled with a definite corresponding diagnosis in order to train and validate the deep learning system.

Reasons for not obtaining photographs included ocular disorders such as severe media opacities (cataracts, cornea scars, etc.), nystagmus and other abnormal eye movements precluding fundus photographs, inability for the patient to sit up, or lack of cooperation (cognitive impairment).

### ***Reference standard (ground truth):***

The deep learning system is trained to recognize features associated with the reference standard (or ground truth). For a diagnostic test, the definition of the reference standard is essential because it serves as the gold standard. This reference standard can be based on diagnoses provided by expert clinicians.<sup>1-4</sup>

Each participating center included patients (and fundus photographs), based on strict inclusion criteria. Centers were asked to provide good quality fundus photographs that included the optic disk, obtained on patients with well-defined neuro-ophthalmic disorders including: 1) “papilledema” (defined as optic disk edema secondary to proven intracranial hypertension, including from an intracranial mass, hydrocephalus or cerebral venous thrombosis, detected on neuroimaging, or secondary to definite idiopathic intracranial hypertension, defined by the modified Dandy criteria with proven elevated cerebrospinal fluid opening pressure on lumbar puncture);<sup>5</sup> 2) “other optic disk abnormalities” including non-glaucomatous optic atrophy of any cause (compressive, ischemic, hereditary, chronic inflammatory optic neuropathies, etc.); optic disk edema secondary to acute anterior optic neuropathies (such as anterior ischemic optic neuropathies and inflammatory anterior optic neuropathies); optic nerve head drusen (buried drusen and non-buried drusen), and congenitally anomalous optic nerves (pseudopapilledema, including tilted optic nerves, small crowded disks, myelinated nerve fibers, etc.).<sup>6,7</sup> Photographs of optic disk drusen were included only if the diagnosis was firmly established by standard imaging criteria, such as B-Ultrasound, optical coherence photography, or fundus autofluorescence photography.<sup>8</sup> Patients with unclear diagnosis, multiple ocular conditions, co-existent retinal conditions or glaucoma were not included. Centers were

encouraged to submit photographs taken with various digital fundus cameras and with various fields, except for ultrawide field fundus cameras which are not appropriate for optic disk photographs.

Each center was also asked to provide photographs of normal appearing optic nerve heads (obtained either in contralateral totally healthy eyes, or in individuals with no past or current history of ophthalmic diseases and no signs of optic neuropathy). Since ocular fundus photographs are not routinely performed in healthy individuals, we also included datasets from three centers of normal subjects with normal optic disks: 1) The center of Healthy Indians, from The Central India Eye and Medical Study, from which we randomly selected 1911 curated normal optic disks;<sup>9</sup> 2) The Singapore Epidemiology of Eye Disease (SEED) Study from which we randomly selected 4053 normal Asian subjects with curated normal optic disks;<sup>10</sup> 3) A group of 330 normal subjects with normal optic disks from France (largely white subjects) [unpublished].

### **a-3. Inclusion and exclusion of photographs (by the Singapore principal site):**

#### ***Non-inclusion and exclusion of images:***

Of a total of 17,470 photographs received from all centers by the Singapore Center, 1,471 photographs were not included in the study because of deficient or incomplete data (see flow charts **Figure S1**), and 153 photographs were excluded because of insufficient quality.

Upon receiving each batch of photographs from each participating center, DM and CV (from the Singapore Center) reviewed all data in detail prior to any analysis and verified that all photographs were correctly matched to a specific patient and that no duplicative images were included for each patient. Additionally, specific attention was paid to the final diagnosis (reference standard or ground truth) provided for each image. In doing so, DM and CV excluded duplicate images (for example, a patient with papilledema who had photographs taken at multiple visits showing similar degrees of papilledema) and those of patients who developed secondary optic atrophy from chronic papilledema and had photographs of optic atrophy provided instead of just papilledema. When the final diagnosis did not seem definite, the images were not included (for example, a patient with bilateral optic disk swelling, characterized by the site principal investigator as “papilledema”, but without available objective evidence of raised intracranial pressure measured by lumbar puncture,<sup>5</sup> or evidence of an intracranial mass by neuroimaging. Another example is the category “optic disk drusen” sometimes diagnosed based on clinical criteria alone, without firm diagnostic evidence on ancillary investigations.<sup>6,8</sup> By applying these stringent inclusion criteria, we secured a valid and robust reference standard, which is of paramount importance for the training of the deep learning system (to avoid the “garbage in, garbage out” problem).

The second group of non-included images was due to data entry mislabelling on the Excel spreadsheet, which made it impossible to match the fundus photograph with the corresponding clinical entry in the Excel template. These images were not usable and were not included.

Finally, 153 photographs (121 in the training datasets and 32 in the external testing datasets) were excluded from the study because of poor image quality that would have resulted in ungradable images, or poor centration of the disk with margins cut at the edge of the photograph. This quality analysis was initially performed manually by CV and DM for a few centers (Atlanta, Angers, Chennai, Freiburg). Soon after, our team developed a procedure able to perform a fully automated quality analysis by a dedicated quality algorithm. This algorithm is able to automatically exclude images with suboptimal quality or incorrect anatomic location resulting in undetectable optic disks prior to automatic cropping of the image and analysis by the diagnostic algorithm (see below in Section S2c, Technical Aspects). Indeed, it is expected that fundus photographs will be non-analysable by the deep learning system if the quality is poor (such as unfocused image, blurry image, major artefacts, eyelash artefacts) or if the image is eccentric, resulting in incomplete visualization of the optic disk. Automated identification of these ungradable or non-analysable photographs by the system prior to analysis is an important step when using any real-life ophthalmic imaging system, as it should prompt repeat fundus photographs and ophthalmology consultation if no photographs of good quality are available. Such quality algorithms are already available and used for the screening of diabetic retinopathy using artificial intelligence in a commercially available fundus camera.<sup>1</sup>

### ***Inclusion of images and demographics***

The remaining 15,846 fundus photographs (7,532 patients, mean (95%CI) age, 48.6 (48.2-49.1) years; age-range, 3 to 98 years; 43.4% men and boys) were used for the training, validation and testing of our deep learning system. The demographic distribution of patients included in the training and external testing datasets is described in **Table S2**. The majority of our patients (5346/7532, 71%) had photographs taken of each optic disk (both eyes); 18% (1327) had a unilateral optic disk photograph, and 11% (859) had multiple photographs during their follow-up, explaining why the total number of patients included in the study is less than the number of photographs analysed.

### ***Uni- or bilaterality of optic nerve diseases:***

The artificial intelligence-deep learning system was trained and tested to detect optic disk abnormalities at the eye level (looking at only one eye) and not on the patient level (looking at one patient's pair of eyes). The results of the optic nerve appearance in the fellow eye was not provided to the machine. Of course, the reference standard clinical diagnosis of each optic nerve disorder took the fellow eye into consideration because the diagnosis was made at the patient level (taking into

account the appearance of both optic nerves for each patient). Some patients had bilateral abnormal optic nerves (such as those with bilateral papilledema or bilateral anterior optic neuritis), whereas some patients had one abnormal optic nerve and one normal optic nerve (such as those with anterior ischemic optic neuropathy or unilateral optic neuritis). For some of these patients, the investigator provided a photograph only of the abnormal optic nerve. Additionally, some patients had an ocular problem in one eye which precluded good quality photographs from being taken (such as when there was a previous ocular trauma, or cataract). This discrepancy has no impact on our study since our algorithm was trained to identify optic nerve abnormalities at the eye level independent of any clinical information and independent of any findings in the fellow eye. In fact, this is an essential strength of our deep learning system because optic nerve abnormalities may be unilateral or bilateral regardless of the underlying pathology.

## **b) List of Digital Retinal Cameras Used in Each Participating Center**

A variety of retinal cameras were used to capture ocular fundus photographs in order to ensure that our deep learning system could be used on photographs obtained with many different digital retinal cameras (see list of cameras in **Table S1**). Wide-field digital cameras were not used.

## **c) Model Development: Technical Aspects**

### **c-1. Definition of artificial intelligence, machine learning and deep learning:**<sup>1-4,11,12</sup>

- i. Artificial Intelligence (AI) refers to a software that can mimic cognitive functions such as learning and problem solving by processing and recognizing patterns in large amounts of data.
- ii. Machine learning creates its own algorithms by “learning” the associations between the input and the output, either in a supervised or un-supervised manner. Supervised learning is defined as training a model with input data and its corresponding labels; unsupervised learning is training a model to identify patterns within the input data without the use of labels.
- iii. Deep learning utilizes multiple processing layers to learn representation of data with multiple levels of abstraction. Deep learning approaches use complete images, and associate the entire image with a diagnostic output, thereby eliminating the use of “hand-engineered” image features.

As is usually done in machine learning, data is split into two major sets: 1) training and validation datasets, and 2) external testing datasets.<sup>1-4</sup> These datasets must not intersect; an image that is in one of the datasets (e.g., training) must not be used in any of the other datasets (e.g., testing). In order to build a robust deep learning system, it is important to have two main components – the ‘dictionary’ (the datasets) and the ‘brain’ (deep neural network – Convolutional Neural Network [CNN]). For training and validation, the deep learning system requires a large training dataset

consisting of images (in our study, fundus photographs showing the optic disk matched with a specific diagnosis provided by expert neuro-ophthalmologists), and selection of a convolutional neural network. Most of the dataset is used for training and validation, followed by external testing, with no overlap of the same data and images in any phases to avoid images from the same patient being used in the training and testing phases. It is preferable to use several independent external testing datasets. The definition of training, validation and external testing datasets are explained below.

**c-2. Definition of training, validation and external testing datasets: <sup>1-4</sup>**

- i. Training dataset: Training of deep neural networks is generally done in batches (subsets) randomly sampled from the training dataset. The training dataset is what is used for optimizing the network weights via backpropagation. Training is performed by updating the model parameters repeatedly until the model optimally fits the data.
- ii. Validation dataset: Validation is used for parameter selection and tuning and is customarily also used to implement stopping conditions for training.
- iii. External testing dataset: Finally, it is important to evaluate the classification performance of the artificial intelligence system using independent datasets, captured using different cameras, populations and clinical settings. This will ensure the generalizability of the system in clinical settings.

**c-3. Definition of a convolutional neural network:**

A convolutional neural network (CNN) is a deep learning algorithm commonly applied to analyse visual data (images).<sup>1-4,13</sup> It can take in an input image, assign importance (learnable weights and loss function) to various aspects/objects in the image and is able to differentiate one aspect/object from another. With training, CNNs have the ability to learn these filters/characteristics. The architecture of a convolutional neural network is analogous to that of the connectivity pattern of neurons in the human brain and resembles the biological processes of the visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. The term “convolutional neural network” indicates that the network employs a mathematical operation called convolution to transform the input volume into an output volume by passing the information from layer to layer similar to what happens in the visual cortex. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms



were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.<sup>1-4,13</sup>

#### **c-4. Development of a deep learning classification model of optic disk abnormalities:**

All technical training of the deep learning system was performed using a NVIDIA Geforce Titan Xp 12GB GDDR5X Graphic Processing Unit (GPU) with Keras and Tensorflow.

##### **Quality control algorithm:**

In our deep learning system, our automated segmentation algorithm first performs a quality check on the ocular fundus photograph to ensure the optic disk is adequately detected. This algorithm is able to detect the optic disk images with insufficient view of the optic disk due to incorrect field of view, or poor image quality secondary to media opacities, eye movement or poor focus. Specifically, this algorithm, trained using multiple convolutional neural networks, analyzes the optic disk size, shape, margins, color balance, exposure, brightness and sharpness, followed by automatic cropping of the optic disk image into a rectangular shape prior to the commencement of the classification task analysis.

##### **Optic disk segmentation:**

A pixel-level semantic image segmentation network (U-Net<sup>14-17</sup>) was used for the localization of the optic disk region. A total of 6370 fundus photographs were labelled with the optic disk region masked at pixel level. The fundus photographs and the masks were used to train the segmentation network. The trained segmentation network was then applied to segment the optic disk region automatically. In order to reduce the number of neural network parameters, a lightweight U-Net inspired by MobileNetV2<sup>18</sup> and U-Net<sup>14</sup> was utilized for the segmentation network, which consisted of an encoder network and a decoder network. The encoder network consisted of 5 convolutional network layers. The decoder network used convolutional transpose for upsampling. A probability density map was produced as the segmentation output, which generated loss gradients for neuron weights updating along the back propagation path.

##### **Optic disk classification:**

Based on the optic disk location, identified by the segmentation network, three optic disk images of different sizes were cropped out for each image. Subsequently, standard data augmentation techniques were used (optic disk rotation, flipping, and random drop out on certain input regions), in order to further increase the number and variety of training samples. Of these three optic disk images,

one optic disk has the same size as the segmentation result and the other two optic disks have a slightly smaller and larger size respectively.

The input optic disk images (224x224 pixels) were trained first using two DenseNet<sup>13,19</sup> (DenseNet-121 and DenseNet-201) pre-trained on ImageNet<sup>20</sup> images. Once the training was done, these two fine-tuned DenseNet were used for feature extractions on optic disk images. At the last convolutional layer of these two fine-tuned DenseNets, the feature vectors were fused into the fully-connected neural network with a softmax layer to optimize the performance. During the training process, optic disk classification output was compared to the reference standard clinical classification provided by expert neuro-ophthalmologists. Discordant findings between the deep learning system and clinical classification was used as error signals to be back propagated, allowing the networks to adapt their neuron weights iteratively in order to reduce the error. This process was repeated for all training images until the network reached a satisfactory performance.

The training started with multiple iterations with a batch size of 32 images, with the initial learning rate of 0.01 and stopped at 60 epochs. For each training iteration, a stochastic gradient descent algorithm was used to optimize a pre-defined loss function to train neuron weights via backpropagation<sup>17,21,22</sup>; at every epoch, the performance of the convolutional neural network (CNN) was assessed using the validation dataset. Based on 5-fold cross-validation results obtained on primary datasets, we adjusted thresholds to achieve sensitivities of at least 90% on the validation datasets.

Subsequently, using the same thresholds, the diagnostic performance of the deep learning system was assessed on 5 independent external testing datasets for detection of the three classes: 1) normal; 2) papilledema; 3) other disk abnormalities. To report performance characteristics for this 3-class classification system, we employ the one-vs-rest strategy<sup>23</sup> (e.g., normal vs all abnormal disks [including papilledema and other disk abnormalities]), and report the AUC, sensitivity, specificity and accuracy for the below 3 cases based on the outputs from our multi-class deep learning system: 1) normal vs abnormal disks (including papilledema and others); 2) papilledema vs non-papilledema (including normal disks and other optic disk abnormalities) and; 3) other disk abnormalities vs normal and papilledema. We also calculated the multi-class AUC for the 5 external testing datasets (see **Table S5**). A multi-class AUC provides an overall average performance of a system for the classification of multiple outcomes (3 outcomes in our study: normal optic disks, papilledema and other disk abnormalities). We computed the overall multi-class AUC by averaging all possible pairwise combinations of classes for each external testing site.<sup>24</sup> The overall performance of our multi-class classification model yielded AUCs over 0.93 that were within the range of the one vs rest strategy.

### **Handling of class imbalance**

To ensure that the training, validation and testing of the system were not affected by variability in the number of cases and controls from each site, we performed the following:

- First, we pooled the data received from 19 sites to generate a deep learning system training dataset with a large sample of optic disk images for each classification. This is a standard procedure in all artificial intelligence/deep learning studies.
- Next, during the deep learning system training process, we further addressed the imbalance between abnormals and normals by performing data augmentation and setting weights to loss function. Data augmentation computationally modifies the input data during the training process to increase the effective data set size and improve both overfitting and data accuracy. We used weighted cross entropy as our loss function during model training with class weights as shown below to handle class imbalance.

$$L = -(w_1 y_1 \log(p_1) + w_2 y_2 \log(p_2) + w_3 y_3 \log(p_3))$$

where  $w_i$  is the class weight for class  $i$ ,  $y_i$  is the label for class  $i$  using one-hot encoding, and  $p_i$  is the predicted probability for class  $i$ . The weight is calculated directly inverse proportional to the number of samples in the corresponding class. It will put more weights to the loss function on the class that has less samples, and less weights on the class that has more samples.

Both methods are also commonly used to correct for frequently occurring class imbalance in deep learning procedures.<sup>25</sup>

- We then conducted a 5-fold cross validation on the training/validation datasets to ensure that the validation results we were observing were not simply obtained by chance. A 5-fold cross validation procedure entails running the experiment 5 times, each time with a random 20% sample acting as a “validation dataset”. This procedure reduces or even eliminates the risk of selection bias. **Figure S3** shows the results obtained on each cross-validation dataset ( $n = 5$ ) displaying the consistent performance of our trained deep learning system.

The differences between the proportions of abnormals and normals in our external testing datasets were expected and are likely to happen in any real-life clinical setting. Considering that our deep learning system displayed a consistently high performance (AUC range: 0.85 to 0.99) in the independent 5 external testing centers, it is unlikely that the accuracy of our system was affected by the proportions of abnormals and normals at each site.

### **Hyperparameter tuning and cross-validation**

In this study we used cross-validation<sup>26,27</sup> to determine the most suitable hyper-parameters for our final model by following the 3 steps below:

**Step 1:** We performed the hyperparameter tuning for the learning rate, optimizer, batch size, and global pooling strategy.

In order to evaluate the performance of each unique set of hyperparameters, we ran 5-fold cross-validation on the training dataset including training 5 separate models using the same set of hyperparameters, followed by the generation of the aggregated performance (area under curve, AUC) of the 5 models. The aggregated performance of the 5 models is used to evaluate the performance of the set of hyperparameters.

Please find below for the range of tuning adopted for each hyperparameter:

- For learning rate, the searching range was 0.001 to 0.05.

- For optimizer, we evaluated stochastic gradient descent (SGD), Adam, RMSprop, Adagrad, Adadelta.

- For batch size, the searching range was from 16 to 48.

- For global pooling strategy, we evaluated global max pooling and the global average pooling.

**Step 2:** Upon completion of the hyperparameter tuning, the final hyperparameters chosen were learning rate = 0.01, optimizer = SGD, batch size = 32, pooling = global average pooling. Using these hyperparameters, we report 5-fold cross-validation performance metrics (AUC, Sensitivity, Specificity and Accuracy) as in **Manuscript Table 2**.

**Step 3:** Using the above-chosen hyperparameters, we re-trained the deep learning system using the entire training dataset and evaluated its diagnostic performance on the testing datasets. Results are reported in **Manuscript Table 2**.

### **Rationale for using a combination of neural models**

We used a method called ensemble learning<sup>28-31</sup> with feature fusion to combine decisions from two networks. Ensemble learning is a widely-used machine learning approach that has been shown to improve classification performance. We have also tested the performance of single neural network models separately (i.e., DenseNet 121 and DenseNet 201). As highlighted in **Table S3**, the combination of DenseNet 121 and DenseNet 201 allowed for a better overall classification performance of the DLS on the external testing dataset compared to a single network.

### **Heatmap Generation**

In order to generate the heatmap, the classification activation map (CAM)<sup>32</sup> was utilized to apply global average pooling on the last convolutional layer in the deep convolutional neural networks. The trained weights for each output from the global average pooling layer indicated the importance/relevance of each feature map from the last convolutional layer. The trained weights were then applied on the corresponding feature maps, which were superimposed on original optic disk images, thus creating class-discriminative visualization in the generated heatmap.

## **d) Statistical and bootstrapping procedures**

### **d-1. Confidence interval estimation:**

Bootstrapping was used only to estimate 95% confidence intervals (CI) for the performance metrics of our classification results (i.e., AUC, sensitivity, specificity and accuracy). We applied n-out-of-n bootstrap with replacement at patient level from our dataset. For each bootstrap sample, we calculated and reserved the performance metrics for that bootstrap sample. The bootstrap sampling was repeated for 2000 times. We then estimated the 95% CI by using the 2.5 and 97.5 percentiles of the empirical distribution of corresponding metrics.

### **d-2. Accuracy and predictive values calculations:**

Accuracy, representing the fraction of correct classifications performed by the deep learning system, was calculated as:

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + False\ negative}$$

Positive and negative predictive values for **Manuscript Table 3** and **Figure S5** below were calculated using the sensitivity and specificity of the deep learning system, after taking into account the prevalence of each condition using the formulas below:

$$PPV = \frac{Sensitivity \times Prevalence}{Sensitivity \times Prevalence + (1 - Specificity) \times (1 - Prevalence)}$$

$$NPV = \frac{Specificity \times (1 - Prevalence)}{Specificity \times (1 - Prevalence) + (1 - Sensitivity) \times Prevalence}$$

where, PPV is the positive predictive value and NPV is the negative predictive value.

## SECTION S3: ADDITIONAL ANALYSES

**a) Errors in Classification by the Deep Learning System**

The deep learning system misclassified a total of 177 fundus photographs in the external testing dataset. These photographs and corresponding clinical information were individually reviewed by 4 expert neuro-ophthalmologists (DM, CV, VB, NJN) in order to understand these discrepancies. **Figures S6A, S6B and S6C** provide details regarding the subgroups of missed normal optic disks (80/613 eyes), missed other optic disk abnormalities (79/532 eyes) and missed papilledema (18/360).

**b) Labelling errors**

The four expert neuro-ophthalmologists met to individually review the 177 photographs that had been misclassified by the deep learning system. They identified 10 labelling errors and these 10 photographs were subsequently relabelled and reclassified after contacting each site PI for clinical confirmation. The results of the re-classification are shown in **Figures S6A, S6B and S6C**. Subsequently, we retested the newly corrected testing dataset which resulted in an improved overall average AUC by a marginal  $0.0077 \pm 0.008$  (from AUC = 0.941 to 0.948).

Given these labelling errors, we subsequently asked each of the 5 external testing centers to individually re-validate the labelling of each of the 1505 photographs included in the study, without providing them the results of the classification obtained from the deep learning system. As a result of this procedure, 3 additional annotation errors (diagnoses assigned) were identified by 2 centers, all within the “other optic disk abnormalities” category. The relabelling of these 3 photographs’ diagnoses did not change their respective classification, which remained “other optic disk abnormalities”, and therefore did not alter our results.

Taking into account all mislabelled images (10 identified by the 4 neuro-ophthalmologists and confirmed by providers + 3 subsequently detected by providers after annotation quality checking), only 0.9% (10/1505) of our testing dataset were mislabelled as regards to one of the 3 categories.

**c) Prevalence of abnormal conditions and predictive values of the deep learning system**

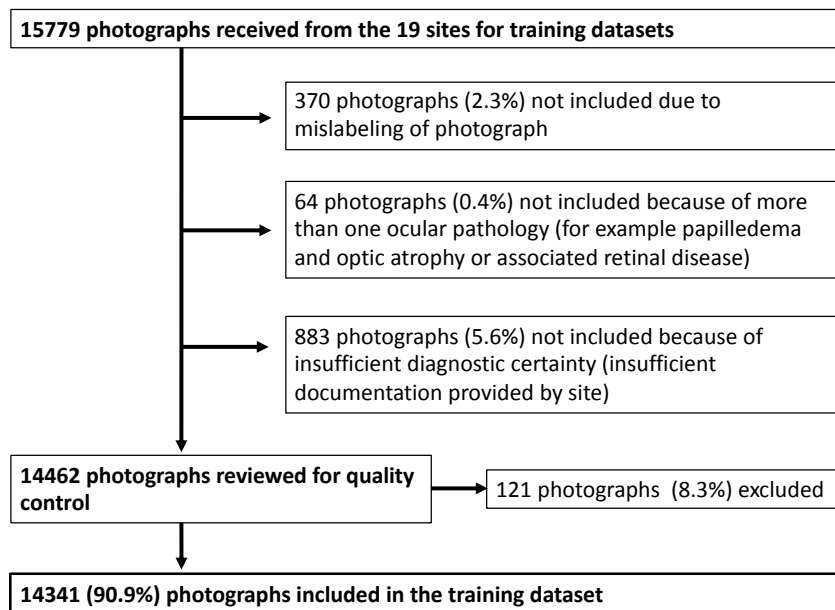
In order to address the influence of prevalence rates of the optic nerve diseases on predictive values, we reached out to our collaborators at each sampled testing site (Bangkok, Copenhagen, Freiburg, Rochester and Teheran) and requested that they provide the prevalence of optic disk

abnormalities and papilledema seen in their respective neuro-ophthalmology clinics (results in **Table S6**). Using these prevalence and the respective performance characteristics of our system, we calculated the predictive values for the classification of papilledema and other optic disk abnormalities at each testing site (**Manuscript Table 3**). We also provided the predictive values of our deep learning system to classify normal optic disks, disks with papilledema, and disks with other optic disk abnormalities across a range of prevalence values (**Figure S5**).

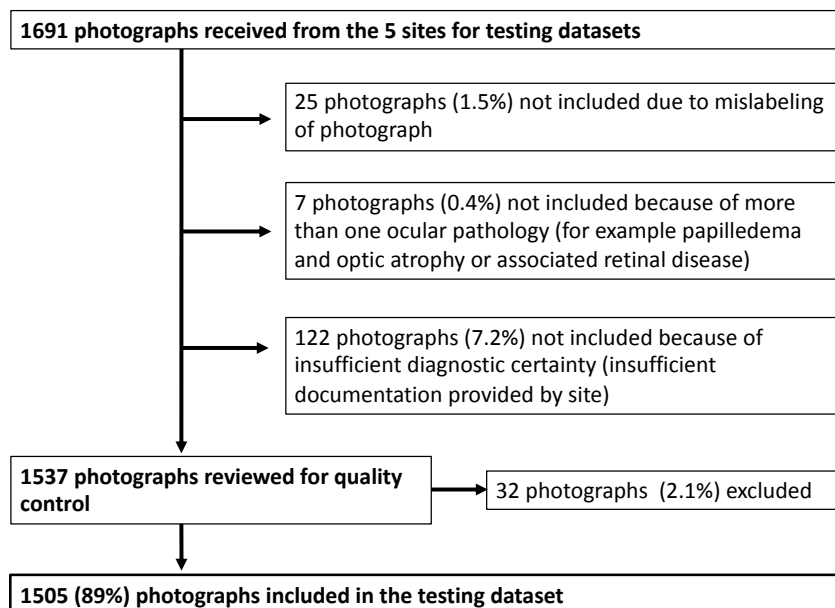
## SECTION S4: FIGURES

**Figure S1:** Flow chart showing the process for inclusion and exclusion of ocular fundus photographs.

**Figure S1A: Training datasets:**



**Figure S1B: External testing datasets:**





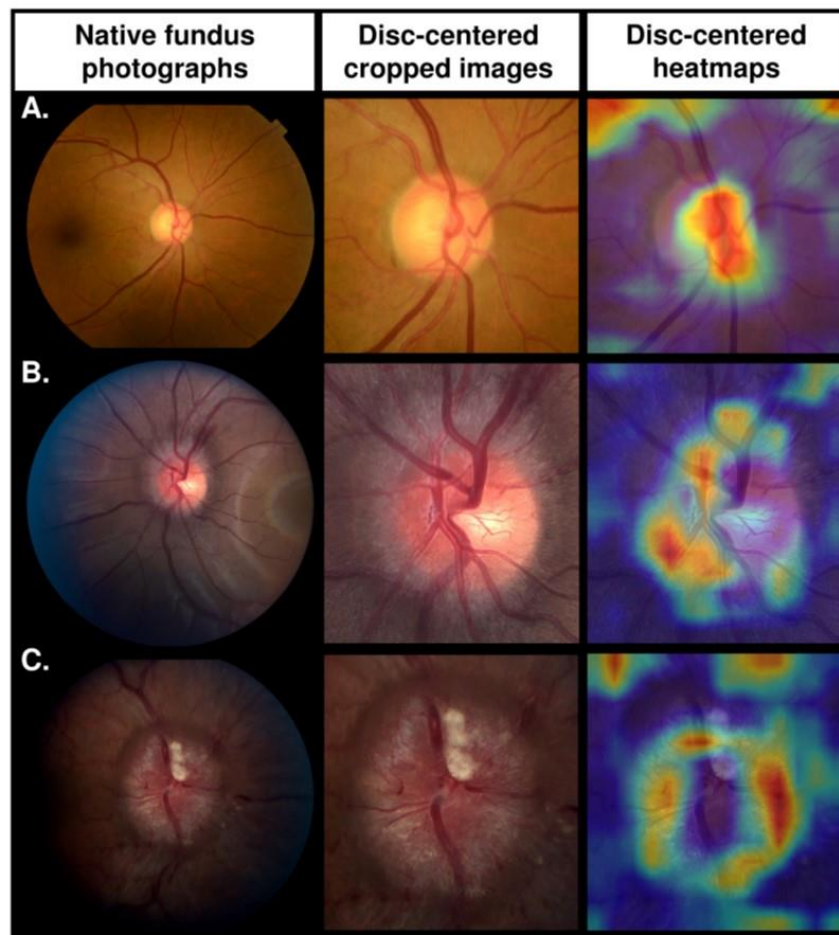
## Figure S2: Examples of fundus photographs and corresponding heatmaps

### Figure S2A: Examples of original (native) fundus photographs, cropped images centered on the optic disk and corresponding class activation maps (heatmaps)

**A-** Normal optic disk in an Asian patient. Diagnostic prediction by the deep learning system: normal 99.99%, papilledema <0.01%, other <0.01%.



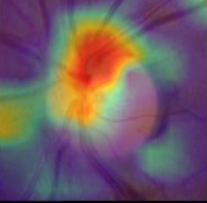

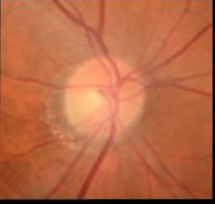
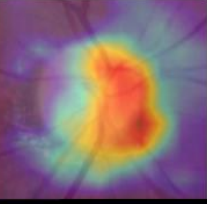

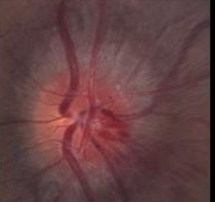
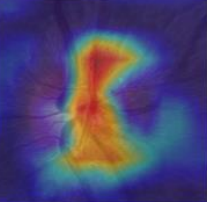


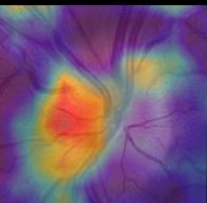

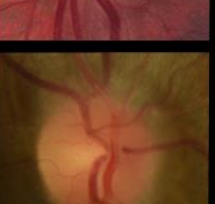
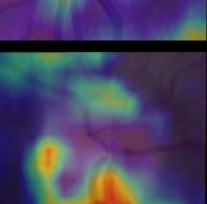
**B-** Mild papilledema in an African-American patient. Diagnostic prediction by the deep learning system: papilledema 99.98%, normal 0.01%, other <0.01%.


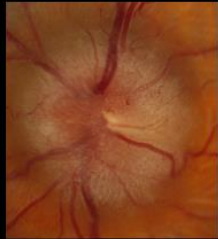
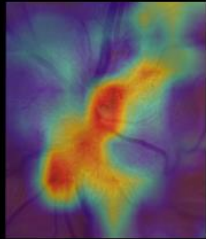


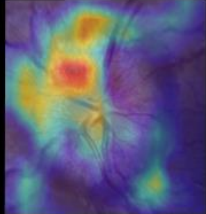


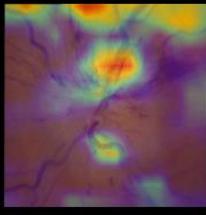

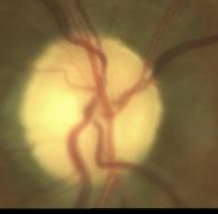
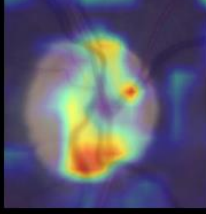

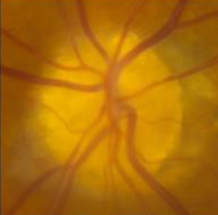
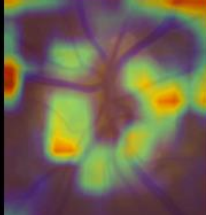
**C-** Severe papilledema in an African-American patient. Diagnostic prediction by the deep learning system: papilledema 99.99%, normal <0.01%, other <0.01%.



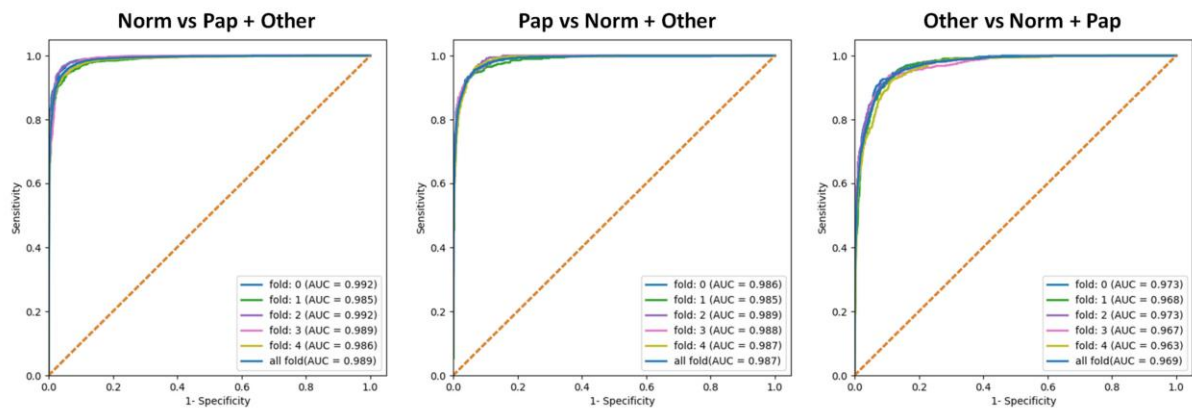
**Figure S2B: Additional examples of fundus photographs and corresponding heatmaps**

Examples of normal and abnormal optic disks with original images as submitted by investigators (left column), cropped images for image analysis by the deep learning system (middle column), and corresponding heatmaps (right column). Photographs were obtained using various digital fundus cameras, with different magnification, on patients of various ethnicities and pigmentation.

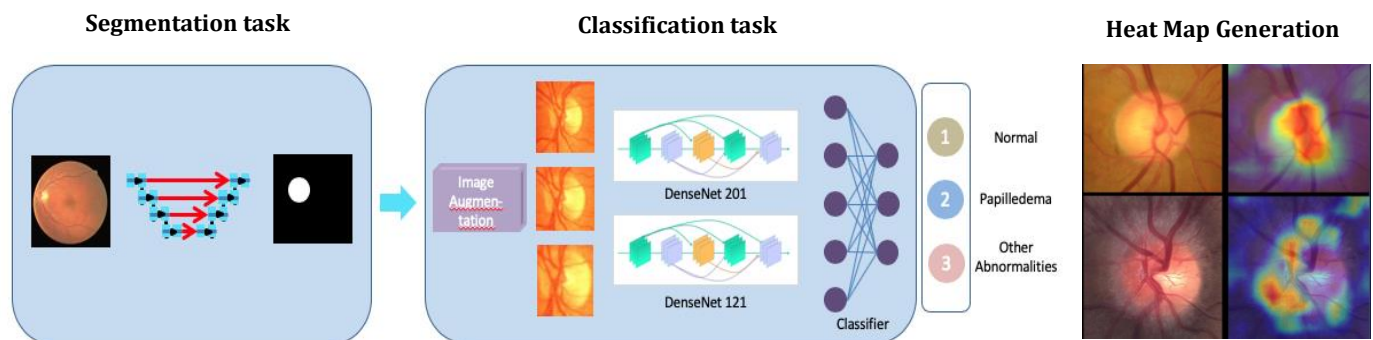
|   |   |   |  |
|---|---|---|--|
|    |    |    | <p><b>Normal optic disc in an Asian patient</b><br/>Diagnostic prediction by the DLS: normal 100%, papilledema &lt;0.01%, other &lt;0.01%.</p>             |
|    |    |    | <p><b>Normal optic disc in an Asian patient</b><br/>Diagnostic prediction by the DLS: normal 99,99%, papilledema &lt;0.01%, other &lt;0.01%.</p>           |
|   |   |   | <p><b>Mild papilledema in an African-American patient</b><br/>Diagnostic prediction by the DLS: papilledema 99,99%, normal &lt;0.01%, other &lt;0.01%.</p> |
|  |  |  | <p><b>Mild papilledema in a white patient</b><br/>Diagnostic prediction by the DLS: papilledema 99,99%, normal &lt;0.01%, other &lt;0.01%.</p>             |
|  |  |  | <p><b>Mild papilledema in an Asian patient</b><br/>Diagnostic prediction by the DLS: papilledema 99,99%, normal &lt;0.01%, other &lt;0.01%.</p>            |

|   |   |   |   |
|---|---|---|---|
|    |    |    | <p><b>Severe papilledema in a white patient</b><br/> Diagnostic prediction by the DLS: papilledema 99.99%,<br/> normal: &lt;0.01%, other: &lt;0.01%</p>                         |
|    |    |    | <p><b>Severe papilledema in an African-American patient</b><br/> Diagnostic prediction by the DLS: papilledema 99.99%,<br/> normal: &lt;0.01%, other: &lt;0.01%</p>             |
|    |    |    | <p><b>Non-arteritic anterior ischemic optic neuropathy in a white patient</b><br/> Diagnostic prediction by the DLS: Other 99.98%,<br/> normal &lt;0.01%, papilledema 0.02%</p> |
|   |   |   | <p><b>Optic atrophy in an Asian patient</b><br/> Diagnostic prediction by the DLS: other 99.99%,<br/> normal 0.01%, papilledema &lt;0.01%</p>                                   |
|  |  |  | <p><b>Optic nerve drusen in a white patient</b><br/> Diagnostic prediction by the DLS: other 99.96%,<br/> normal 0.03%, other &lt;0.01%</p>                                     |

**Figure S3:** Receiver operating characteristic curves and areas under the curves (AUC) of individual folds for the 5-fold cross-validation performed on the primary data-set.

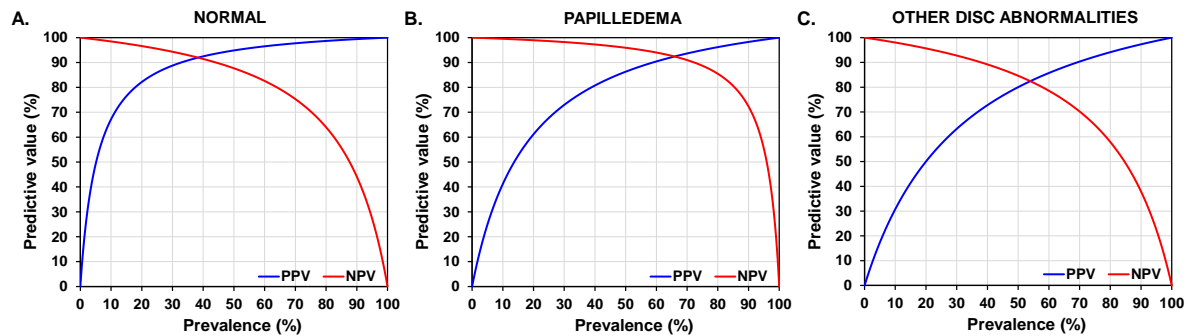


**Figure S4:** Technical model with segmentation and classification networks



**Figure S5: Predictive values of the deep learning system across a full prevalence range for the detection of normal disks, disks with papilledema and other optic disk abnormalities**

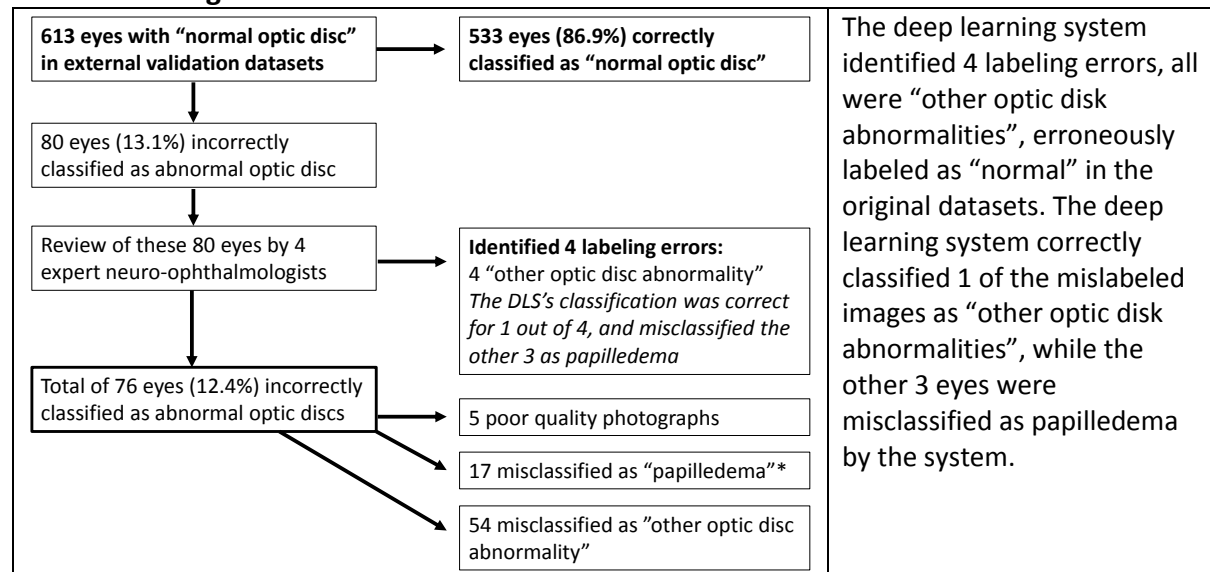
Predictive values of the deep learning system for the detection of normal disks (A), papilledema (B) and other optic disk abnormalities (C) across a full (0 – 100%) range of prevalence. Positive and negative predictive values were derived from the overall performance characteristics (sensitivity, specificity) of the deep learning system in the 5 external testing datasets originating from 5 different neuro-ophthalmology clinics. This figure does not include uncertainty estimates.



**Abbreviations:** PPV: positive predictive value; NPV: negative predictive value

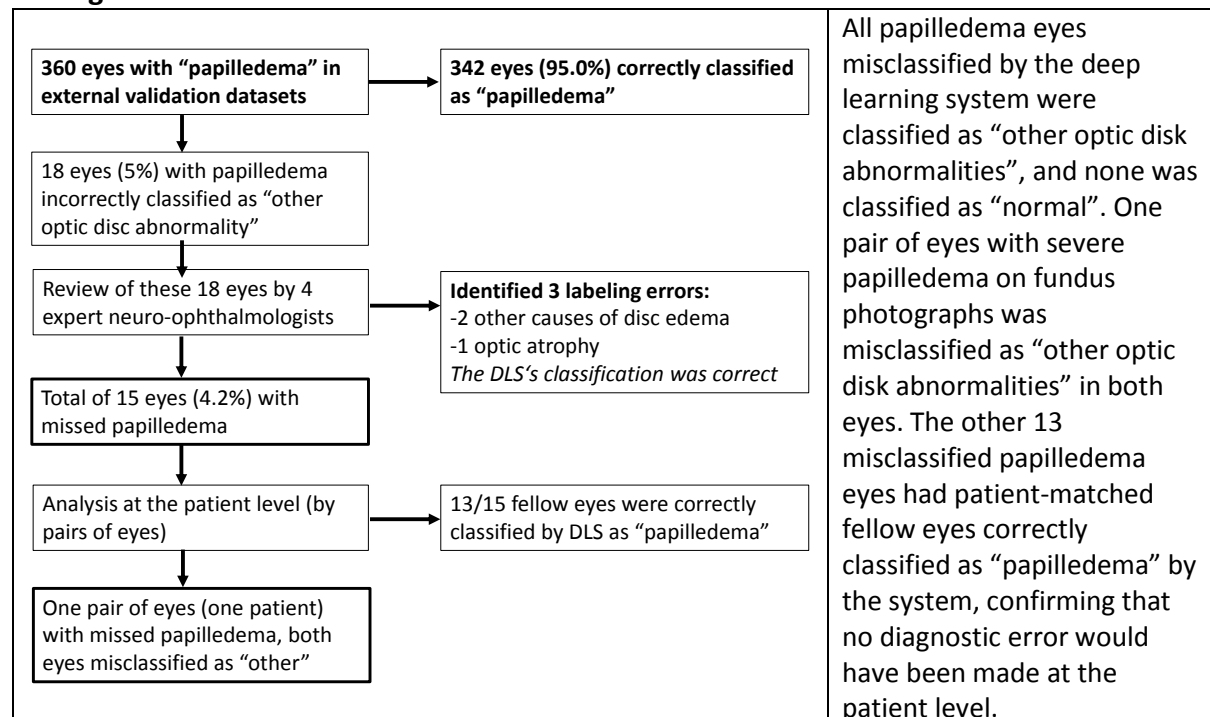
**Figure S6: Errors in classification by the deep learning system**

**Figure S6A: Misclassification of “normal optic disk” by the deep learning system in the external testing datasets.**



\*Among the 17 patients misclassified as “papilledema” by the system, 7 had small crowded optic disks and would likely have been misinterpreted as having mild optic disk edema on fundus photographs, even by expert neuro-ophthalmologists without clinical information.

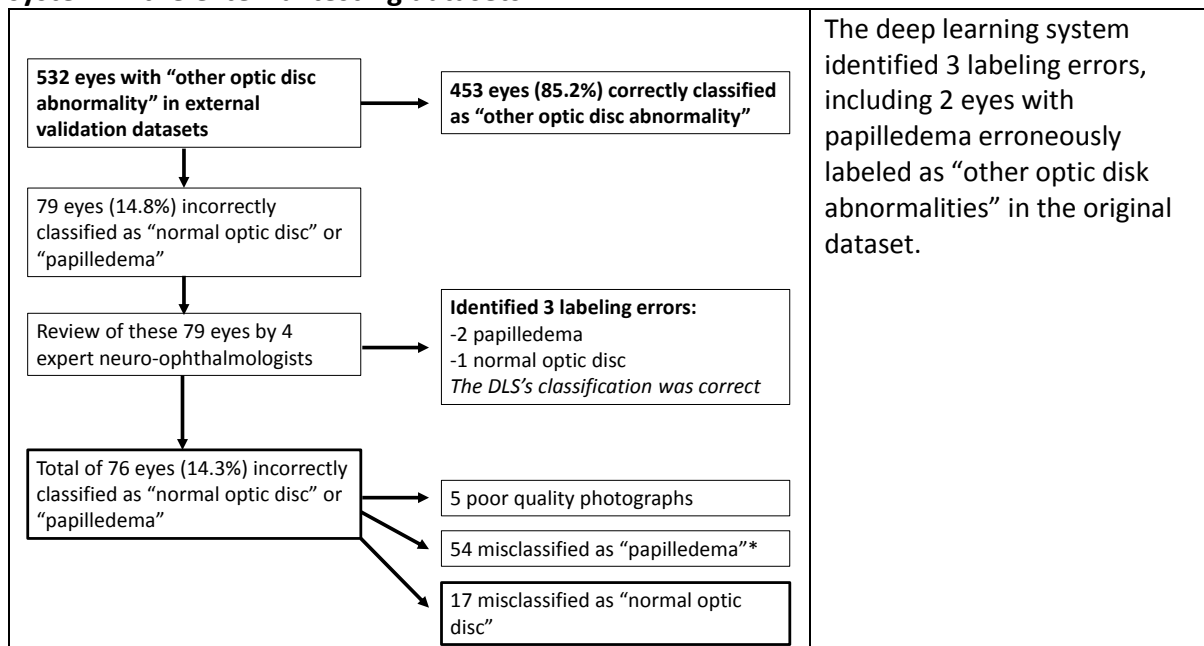
**Figure S6B: Misclassification of “papilledema” by the deep learning system in the external testing datasets.**



The 15/360 (4.2%) photographs with missed papilledema originated from Bangkok (2), Copenhagen (1), Freiburg (2), Rochester (3) and Teheran (7).



**Figure S6C: Misclassification of “other optic disk abnormalities” by the deep learning system in the external testing datasets.**



\*Among the 54 patients misclassified as “papilledema” by the system, 49 had either disk edema from another cause (e.g., anterior ischemic optic neuropathy or anterior optic neuritis) or congenital optic disk anomalies mimicking disk edema and would easily have been misinterpreted as having papilledema on fundus photographs, even by expert neuro-ophthalmologists without clinical information.

Seven eyes with “other optic disk abnormalities” were classified as “normal” by the system, but their appearance was definitely abnormal on fundus photographs.

|                           |
|---------------------------|
| <b>SECTION S5: TABLES</b> |
|---------------------------|

**Table S1:** List of digital retinal cameras used in each participating center

| City, Country        | Camera brand  | Model                      |
|----------------------|---------------|----------------------------|
| Angers, France       | Topcon        | TRC-NW6S                   |
| Atlanta, USA         | Topcon        | 50DX                       |
| Baltimore, USA       | Zeiss         | FF4                        |
| Bordeaux, France     | Zeiss         | VISUCAM                    |
| Bangkok, Thailand    | Kowa          | WX3D                       |
| Bologna, Italy       | Zeiss/Topcon  | VISUCAM 500/DRI OCT Triton |
| Central India        | Zeiss         | FF450                      |
| Coimbra, Portugal    | Topcon        | TRC-NW7SF Mark II          |
| Copenhagen, Denmark  | Topcon        | TRC 50DX and TRC NW8       |
| Chennai, India       | Zeiss         | FF450 Plus IR              |
| Freiburg, Germany    | Zeiss         | SF 420                     |
| Lille, France        | Nidek         | AFC330                     |
| Geneva, Switzerland  | Zeiss         | FF450 Plus                 |
| Grenoble, France     | Topcon/Canon  | TRC NW6S/CR2               |
| Guangzhou, China     | Topcon/Zeiss  | TRC 50DX/FF450 Plus        |
| Hong Kong, China     | Topcon        | TRC 50DX                   |
| London, UK           | Topcon/Canon  | TRC 50DX/CR2               |
| Manila, Philippines  | Zeiss/Meditec | VISUCAM 500/NMFA           |
| Paris, France        | Canon         | CRDGI                      |
| Rochester, USA       | Topcon        | TRC 50DX                   |
| Sydney, Australia    | Zeiss         | VISUCAM 500                |
| Syracuse, USA        | Topcon/Zeiss  | TRC NW8/TRC NW400/FF 450   |
| Singapore, Singapore | Topcon/Canon  | TRC 50DX/CR-Dgi            |
| Teheran, Iran        | Canon         | CR2                        |



**Table S2.** Demographic distribution of patients from the training and external testing datasets.

| City, Country                                   | Number of patients | Age (95% CI), years       | Available age data, % | Gender, % male | Available gender data, % |
|---|--------------------|---------------------------|-----------------------|----------------|--------------------------|
| <b>Primary training and validation datasets</b> |                    |                           |                       |                |                          |
| Angers, France                                  | 492                | 52.1 (50.0 - 54.2)        | 100                   | 44.1           | 100                      |
| Atlanta, USA                                    | 934                | 37.3 (36.2 - 38.4)        | 100                   | 21.9           | 100                      |
| Baltimore, USA                                  | 391                | 30.3 (26.5 - 34.1)        | 10.2                  | 30.8           | 10.0                     |
| Bologna, Italy                                  | 152                | 37.1 (34.6 - 39.6)        | 100                   | 65.8           | 100                      |
| Bordeaux, France                                | 36                 | 51.6 (45.4 - 57.8)        | 100                   | 52.9           | 94.4                     |
| Chennai, India                                  | 338                | 37.6 (36.0 - 39.1)        | 100                   | 46.7           | 100                      |
| Coimbra, Portugal                               | 173                | 53.3 (50.5 - 56.1)        | 100                   | 49.1           | 100                      |
| Geneva, Switzerland                             | 71                 | 42.8 (38.0 - 47.6)        | 100                   | 64.8           | 100                      |
| Grenoble, France                                | 171                | 45.9 (42.2 - 49.5)        | 98.8                  | 50.9           | 98.8                     |
| Guangzhou, China                                | 58                 | 54.1 (51.6 - 56.7)        | 100                   | 55.2           | 100                      |
| Hong Kong, China                                | 394                | 58.2 (56.7 - 59.6)        | 100                   | 37.8           | 100                      |
| Lille, France                                   | 198                | 50.2 (47.9 - 52.6)        | 100                   | 37.4           | 100                      |
| London, UK                                      | 190                | 53 (50.0 - 56.1)          | 100                   | 52.6           | 100                      |
| Manila, Philippines                             | 36                 | 43.8 (37.7 - 49.8)        | 100                   | 61.1           | 100                      |
| Nagpur, India                                   | 521                | 46.6 (45.5 - 47.6)        | 100                   | 49.4           | 99.4                     |
| Paris, France                                   | 138                | 44.1 (41.5 - 46.7)        | 100                   | 44.2           | 100                      |
| Singapore, Singapore                            | 2194               | 56.2 (55.8 - 56.7)        | 99.8                  | 49.2           | 99.8                     |
| Sydney, Australia                               | 259                | 42.2 (39.9 - 44.4)        | 94.6                  | 44.5           | 94.6                     |
| Syracuse, USA                                   | 33                 | 46.5 (39.4 - 53.7)        | 100                   | 30.3           | 100                      |
| <b>External testing datasets</b>                |                    |                           |                       |                |                          |
| Bangkok, Thailand                               | 159                | 49.8 (47.1 - 52.5)        | 99.4                  | 38.0           | 99.4                     |
| Copenhagen, Denmark                             | 101                | 39.7 (36.4 - 43.0)        | 100                   | 25.7           | 100                      |
| Freiburg, Germany                               | 156                | 40.9 (37.4 - 44.5)        | 100                   | 35.3           | 100                      |
| Rochester, USA                                  | 148                | 47.9 (44.9 - 51.0)        | 100                   | 35.4           | 99.3                     |
| Teheran, Iran                                   | 189                | 42.6 (40.4 - 44.8)        | 99.5                  | 48.9           | 99.5                     |
| <b>All centers</b>                              | <b>7532</b>        | <b>48.6 (48.2 - 49.1)</b> | <b>95.0</b>           | <b>43.4</b>    | <b>94.9</b>              |

Some centers had minimal missing demographic data that could not be retrieved. The Baltimore center provided a completely de-identified convenience sample with accurate diagnoses but most demographic data could not be retrieved.

**Table S3: Classification performance of single neural networks against a combination of two networks**

Table comparing the classification performance of single neural networks (DenseNet 121, DenseNet 201) against a combination of the two networks. The combination of 2 networks yielded a higher classification performance, represented here by AUCs, compared to a single network approach.

|                     | Average AUC over all external testing centers |             |               |
|---------------------|---|-------------|---------------|
|                     | DenseNet121                                   | DenseNet201 | Both Networks |
| Norm vs Pap + Other | 0.97  | 0.97        | <b>0.98</b>   |
| Pap vs Other + Norm | 0.95  | 0.95        | <b>0.96</b>   |
| Other vs Norm + Pap | 0.88  | 0.89        | <b>0.90</b>   |

**Table S4:** Classification performance of the deep learning system on the individual external testing datasets.

| One-vs rest classification | City, Country       | Total No. | Normal | Papilledema | Others | AUC (95% CI)       | Sensitivity (95% CI), % | Specificity (95% CI), % | Accuracy (95% CI), % |
|----------------------------|---------------------|-----------|--------|-------------|--------|--------------------|-------------------------|-------------------------|----------------------|
| Norm vs Pap + Other        | Bangkok, Thailand   | 319       | 177    | 38          | 104    | 0.98 (0.97 - 0.99) | 94.9 (88.1 - 97.3)      | 90.8 (84.4 - 96.3)      | 93.1 (88.2 - 95.3)   |
| Pap vs Other + Norm        | Bangkok, Thailand   | 319       | 177    | 38          | 104    | 0.96 (0.94 - 0.98) | 94.7 (87.0 - 100)       | 84.3 (81.3 - 90.2)      | 85.6 (82.9 - 91.0)   |
| Other vs Norm + Pap        | Bangkok, Thailand   | 319       | 177    | 38          | 104    | 0.91 (0.87 - 0.94) | 81.7 (73.3 - 89.7)      | 83.7 (78.4 - 89.4)      | 83.1 (79.0 - 87.9)   |
| Norm vs Pap + Other        | Copenhagen, Denmark | 200       | 90     | 47          | 63     | 0.96 (0.93 - 0.99) | 81.1 (71.6 - 89.5)      | 97.3 (93.8 - 100)       | 90.0 (85.1 - 94.1)   |
| Pap vs Other + Norm        | Copenhagen, Denmark | 200       | 90     | 47          | 63     | 0.98 (0.96 - 0.99) | 100 (100 - 100)         | 89.5 (82.4 - 94.8)      | 92.0 (86.5 - 96.0)   |
| Other vs Norm + Pap        | Copenhagen, Denmark | 200       | 90     | 47          | 63     | 0.91 (0.86 - 0.96) | 92.1 (85.5 - 98.2)      | 69.3 (60.0 - 77.2)      | 76.5 (69.0 - 82.8)   |
| Norm vs Pap + Other        | Freiburg, Germany   | 328       | 98     | 92          | 138    | 0.99 (0.98 - 1)    | 90.8 (85.0 - 96.6)      | 96.1 (92.7 - 98.4)      | 94.5 (91.7 - 97.0)   |
| Pap vs Other + Norm        | Freiburg, Germany   | 328       | 98     | 92          | 138    | 0.96 (0.94 - 0.98) | 98.9 (96.5 - 100)       | 79.2 (72.4 - 85.0)      | 84.8 (79.3 - 89.0)   |
| Other vs Norm + Pap        | Freiburg, Germany   | 328       | 98     | 92          | 138    | 0.92 (0.89 - 0.96) | 87 (80.5 - 93.1)        | 84.2 (78.3 - 89.4)      | 85.4 (81.1 - 89.5)   |
| Norm vs Pap + Other        | Rochester, USA      | 284       | 92     | 95          | 97     | 0.96 (0.94 - 0.98) | 80.4 (71.9 - 89.3)      | 95.8 (92.6 - 98.4)      | 90.8 (87.0 - 94.4)   |
| Pap vs Other + Norm        | Rochester, USA      | 284       | 92     | 95          | 97     | 0.98 (0.96 - 0.99) | 96.8 (93.0 - 100)       | 84.1 (78.7 - 89.8)      | 88.4 (84.5 - 92.5)   |

|                     |                |     |     |    |     |                    |                    |                    |                    |
|---------------------|----------------|-----|-----|----|-----|--------------------|--------------------|--------------------|--------------------|
| Other vs Norm + Pap | Rochester, USA | 284 | 92  | 95 | 97  | 0.94 (0.91 - 0.97) | 94.8 (90.0 - 98.9) | 72.2 (65.6 - 80.3) | 79.9 (75.1 - 85.7) |
| Norm vs Pap + Other | Teheran, Iran  | 374 | 156 | 88 | 130 | 0.98 (0.96 - 0.99) | 87.8 (82.5 - 93.3) | 95.9 (93.0 - 98.6) | 92.5 (89.7 - 95.4) |
| Pap vs Other + Norm | Teheran, Iran  | 374 | 156 | 88 | 130 | 0.93 (0.90 - 0.96) | 92.0 (85.8 - 97.9) | 83.6 (78.5 - 88.0) | 85.6 (81.4 - 89.3) |
| Other vs Norm + Pap | Teheran, Iran  | 374 | 156 | 88 | 130 | 0.85 (0.79 - 0.89) | 77.7 (69.2 - 85.2) | 76.2 (69.0 - 82.3) | 76.7 (71.4 - 81.4) |

The average age of patients included in the external testing dataset was 44.4 years (95%CI: 43.1 – 45.8), based on 99.7% of available patient demographics. The male to female ratio in the testing dataset was 0.61 (38.0% men and boys), based on 99.6% of available patient demographics.

Abbreviations: AUC: area under the receiver operating characteristic curve; CI: confidence interval; Norm: Normal disks; Pap: Disks with papilledema; Other: all other optic disk abnormalities, including non-arteritic anterior ischemic optic neuropathy, optic atrophy, other causes of optic disk swelling, optic disk drusen, congenital optic disk abnormalities, etc. No: number

**Table S5: Multi-class AUC for the 5 external testing datasets and comparison with the range of AUCs for one-vs-rest strategy**

| External test set   | Range of AUCs for one-vs-rest strategy* | Multi-class AUC** |
|---------------------|---|-------------------|
| Bangkok, Thailand   | 0.91 – 0.98                             | 0.95              |
| Copenhagen, Denmark | 0.91 – 0.98                             | 0.95              |
| Freiburg, Germany   | 0.92 – 0.99                             | 0.96              |
| Rochester, USA      | 0.94 – 0.98                             | 0.96              |
| Teheran, Iran       | 0.85 – 0.98                             | 0.92              |
| <b>All centers</b>  | <b>0.90 – 0.98</b>                      | <b>0.95</b>       |

\*normal vs [papilledema and other disk abnormalities]; papilledema vs [normal and other disk abnormalities]; other disk abnormalities vs [papilledema and normal].

\*\*Calculated according to Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 2001;45(2):171–186.<sup>24</sup>

**Table S6:** Estimated prevalences of abnormal conditions at testing sites

| Center,<br>Country         | Estimated prevalence        |                    |                                       | Total number<br>of patients<br>per year |
|----------------------------|-----------------------------|--------------------|---------------------------------------|---|
|                            | Abnormal optic<br>disks (n) | Papilledema (n)    | Other optic disk<br>abnormalities (n) |   |
| Bangkok,<br>Thailand       | 72.2% (325)                 | 8.9% (40)          | 63.3% (285)                           | 450                                     |
| Copenhagen,<br>Denmark     | 17.9% (500)                 | 3.6% (100)         | 14.3% (400)                           | 2800                                    |
| Freiburg,<br>Germany       | 50% (2000)                  | 10% (400)          | 40.0% (1600)                          | 4000                                    |
| Rochester,<br>USA          | 50% (800)                   | 17.2% (275)        | 32.8% (525)                           | 1600                                    |
| Teheran,<br>Iran           | 40% (200)                   | 8% (40)            | 32.0% (160)                           | 500                                     |
| <b>Average<br/>(range)</b> | 46.0% (17.9 – 72.2%)        | 9.5% (3.6 – 17.2%) | 36.5% (14.3 – 63.3%)                  |   |

|   |
|---|
| <b>SECTION S6: REFERENCES FOR SUPPLEMENTARY APPENDIX:</b> |
|---|

1. Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, Schmetterer L, Pasquale LR, Bressler NM, Webster DR, Abramoff M, Wong TY. Deep learning in ophthalmology: The technical and clinical considerations. *Prog Retin Eye Res.* 2019; 72: 100759.
2. Ting DSW, Lee AY, Wong TY. An ophthalmologist's guide to deciphering studies in artificial intelligence. *Ophthalmology.* 2019; 126: 1475-1479.
3. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA.* 2019; 322: 1806-1816.
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019; 380: 1347-1358.
5. Friedman DI, Liu GT, Digre KB. Revised diagnostic criteria for the pseudotumor cerebri syndrome in adults and children. *Neurology* 2013; 81: 1159-65.
6. Biousse V, Newman NJ. Diagnosis and clinical features of common optic neuropathies. *Lancet Neurol* 2016; 15: 1355-67.
7. Biousse V, Newman NJ. Ischemic optic neuropathies. *N Engl J Med* 2015; 372: 2428-36. Erratum in: *N Engl J Med* 2015; 373: 2390.
8. Gise R, Gaier ED, Heidary G. Diagnosis and imaging of optic nerve head drusen. *Semin Ophthalmol* 2019; 34: 256-63.
9. Nangia V, Matin A, Bhojwani K, Kulkarni M, Yadav M, Jonas JB. Optic disc size in a population-based study in central India: The Central India Eye and Medical Study (CIEMS). *Acta Ophthalmol* 2008; 86: 103-4.
10. Cheung N, Teo K, Zhao W, Wang JJ, Neelam K, Tan NYQ, Mitchell P, Cheng CY, Wong TY. Prevalence and associations of retinal emboli with ethnicity, stroke, and renal disease in a multiethnic asian population: The Singapore Epidemiology of Eye Disease Study. *JAMA Ophthalmol* 2017; 135: 1023-8
11. Webb, S. Deep learning for biology. *Nature* 2018; 554: 555–7.
12. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol* 2019 Sep 12. doi: 10.1001/jamaophthalmol.2019.3501. [Epub ahead of print] PubMed PMID: 31513266; PubMed Central PMCID: PMC6743057.
13. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks, Conference on Computer Vision and Pattern Recognition, CVPR 2017.
14. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351.* 2015.

15. Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 Lecture Notes in Computer Science, vol. 9901. 2016.
16. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods* 2019; 16: 67–70.
17. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. *Radiology* 2018; 288: 177-85.
18. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks, *CVPR* 2018.
19. Zhou Q, Zhou Z, Chen C, et al. Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. *Computers in Biology and Medicine* 2019; 107: 47-57.
20. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database, *Conference on Computer Vision and Pattern Recognition, CVPR* 2009.
21. Aggarwal SLP. Data augmentation in dermatology image recognition using machine learning, *Skin Research Technology* 2019 May 29.
22. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521, 436–44.
23. Rifkin R, Klautau A. In defense of one-vs-all classification. *J Mach Learn Res.* 2004; 5: 101–141.
24. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 2001; 45: 171–86
25. Mateusz B, Atsuto M, and Maciej AM. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018; 106:249–259.
26. Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning : with applications in R. New York: Springer 2013.
27. Po-Hsuan CC, Yun L, Lily P. How to develop machine learning models for healthcare? *Nature Materials* 2019; 18: 410–414.
28. Yang P, Yang YH, Zhou BB, Zomaya AY. A Review of ensemble methods in bioinformatics; Including stability of feature selection and ensemble feature selection methods. *Current Bioinformatics.* 2010; 5: 296-308. Available at (accessed on 12/16/2019): <http://www.maths.usyd.edu.au/u/pengyi/publication/EnsembleBioinformatics-v6.pdf>.
29. Boström H. Feature vs. classifier fusion for predictive data mining a case study in pesticide classification. 2007 10th International Conference on Information Fusion, Quebec, Que., 2007, pp. 1-7.
30. Hansen LK, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1990; Vol 12, No. 10.



31. Ashmita S, Shukla KK. Review on the architecture, algorithm and fusion strategies in ensemble learning. Internal J Computer Application. 2014; Vol 108, No. 8.
32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. CVPR 2016: 2921-9.