Eye (2020) 34:1947–1949
https://doi.org/10.1038/s41433-020-0929-6

The ROYAL COLLEGE of OPHTHALMOLOGISTS

EDITORIAL

AI papers in ophthalmology made simple

Sohee Jeon[1] · Yun Liu[2] · Ji-Peng Olivia Li[3] · Dale Webster[2] · Lily Peng[2] · Daniel Ting[4]

Received: 17 March 2020 / Revised: 13 April 2020 / Accepted: 22 April 2020 / Published online: 7 May 2020
© The Royal College of Ophthalmologists 2020

Recently, *EYE* has published few manuscripts on artificial intelligence (AI) systems based on deep learning (DL) [1, 2]. In ophthalmology, with the exponential growth in computational power, ocular imaging quality, and increasing capabilities, several groups have applied AI productively to interpret ocular images for diagnosis, referral management, risk stratification, and prognostication [3–6]. Clinical implementation has also begun with the first FDA-cleared AI-equipped fundus camera for DR screening in 2018 (IDx-DR; IDx Technologies Inc, Coralville, IA, USA).

Many general ophthalmologists may not have a computer science background, and traditional critical analysis skills for clinical studies do not always directly apply to AI studies. This editorial outlines a stepwise approach to help readers critically read the introduction, methods, results, and discussion components of an AI paper, with a view towards how these technologies can potentially be applied in routine clinical practice.

The introduction of the manuscript should describe unmet clinical needs, unique features of the AI system, and how the AI system described aims to fulfill said clinical needs.

In the methodology, the datasets and the AI method, such as an artificial neural network, are the two main components required for AI-based medical image analysis. Three main aspects need to be evaluated here: the input to the AI, the way the AI processes the input, and the AI outputs.

(1) Input to the AI

Input data could be clinical data, medical images, genomics, or all of the above, depending on the specific research question. Fundus photography and optical coherence tomography (OCT) are the most commonly studied ocular images. If the AI system is developed to process fundus photography, the laterality, fields of view (30, 45, 50° or a wide field), and views of the field (optic disc or macula-centered) should be stated. For OCT-based AI studies, the site of the eyes, direction of the scan (vertical or horizontal), and the number of included scans (subfoveal only, central 11 cuts, or total slabs) should be stated.

Is there a minimum size required for the training dataset? Similar to a statistical analysis in a clinical trial, AI performance increases based on the volume of training data, with some variability based on the AI's "task", quality of the images, and ground truth. Thus, there is no one-answer-fits-all answer for the quantity of data needed. Several studies have reported ways to determine sample size needed to develop AI models, such as estimation using a learning curve [7]. Methods to determine the sample size needed for the evaluation of AI include well-studied statistical power calculations for the evaluation metric (such as sensitivity or specificity).

(2) AI processing

DL describes multilayered artificial neural networks, which are inspired by the human brain. The most common type of DL method used for medical images is the convolutional neural network (CNN). Several types of layers that can be found in a CNN include convolutional layers, pooling layers, fully connected layers, and normalization layers. Each artificial neuron contains a weight that is multiplied with its input and frequently "thresholded" to determine whether it should "fire". These weights are automatically learnt during the process of training by exposing the network to many examples. Therefore, training a DL model does not require the manual crafting of predictive features that is needed for ML. To reduce the need for large labeled datasets, transfer learning may be used to help the AI learn from other image interpretation tasks instead of starting from scratch.

Although AI systems learn from examples during the training process, several settings need to be prespecified before training. These settings are called "hyperparameters" [8], and include the learning rate, batch size, momentum, and

✉ Daniel Ting
daniel.ting45@gmail.com

[1] Keye Eye Center, Seoul, Korea
[2] Google Health, Palo Alto, CA, USA
[3] Moorfields Eye Hospital, London, UK
[4] Singapore National Eye Center, Singapore, Singapore

weight decay. Many of these hyperparameters define how the AI system learns from the data. Because these hyperparameters require extensive trial and error, a separate "validation/tuning" dataset is needed to fine tune these hyperparameters.

To accomplish the above, the collected datasets are generally split into train, validation, and test sets. Because of terminology differences between fields, the "validation set" is sometimes called the tuning set; and the test set is sometimes called the "clinical validation set" or "holdout set". Some common options for splitting datasets are 6/2/2, 7/2/1, or 8/1/1, though these ratios are empirical. Cross-validation can also be used to minimize the sampling bias from splitting of datasets, though its use in development (multiple training/validation splits) versus evaluation (multiple testing splits) has different goals and beyond the scope of this article.

What may come as a surprise to some readers new to DL is the complexity of the algorithm that processes the input to produce the output. Though the exact weights and multiplications used are known precisely, there are typically millions or billions of such mathematical operations, which makes understanding it difficult. As such, some have labeled such methods a "black box". Clinicians may understandably be reluctant to accept research outcomes and clinical decisions made through algorithms that are not completely understood. Furthermore, the real-world implications are that clinicians may still retain medico-legal accountability when using AI in clinical decision making [9]. Thankfully, several methods exist to understand how DL-based AI interpretes images. One of the most commonly used methods is the occlusion-based procedure, in which an algorithm is repeatedly tested with parts of the input occluded to create a map showing which parts of the data influenced the output [10, 11].

(3) Output of the AI

AI performance is evaluated by comparison with a reference standard, which is often a widely accepted gold standard or ground truth. The reference standard is vital in validating an algorithm, and is often based on the agreement of several professionals, consultant ophthalmologists, fellowship-trained subspecialists, certified nonmedical professional graders, or optometrists in reading centers who have undertaken intensive training and accreditation with reproducible and consistent outcomes.

The result section quantifies the performance of the AI system by reporting the measures of discrimination, such as the area under the curve (AUC), sensitivity (also called true positive rate or recall), specificity (equivalent to 1—false positive rate), positive predictive value (PPV), and negative predictive value (NPV):

(1) AUC: AI performance can be described by receiver operating characteristic (ROC) curves and precision–recall curves. ROC curves summarize the trade-off between the true positive rate (y-axis) and the false positive rate (x-axis) using different thresholds, whereas precision–recall curves summarize the trade-off between the PPV (y-axis) and the true positive rate (x-axis). The AUC measures the entire two-dimensional area underneath the ROC curve; hence, providing an aggregate measure of performance across all possible classification thresholds. An algorithm with a 100% accuracy compared with the "reference standard" has an AUC value of 1.0.
(2) Sensitivity and specificity: The rationale for how an operating threshold is determined during the training phase should be described and sensitivity and specificity should be demonstrated on the independent datasets performed at the same operating threshold. Sensitivity and specificity are generally not influenced by the prevalence of the disease if there is no spectrum bias.
(3) Predictive values: PPV is the proportion of cases that truly have the target condition among cases with positive test results. NPV is the proportion of cases that truly do not have the target condition among cases with negative test results. Unlike sensitivity and specificity, predictive values are affected by the prevalence of the disease. PPV increases while NPV decreases as the prevalence increases.
(4) Interrater comparisons, such as Cohen's k value, are used to evaluate whether there is an agreement between interpretations. In an AI study, it is usually used to describe the interrater agreement between an algorithm and reference standard or human graders. It does not reflect the true accuracy of the test.
(5) The diagnostic accuracy or the effectiveness of a test, is the ability to discriminate between the target condition and health correctly in binary classification. It is calculated by the proportion of correct predictions (including both positive and negative predictions) among all evaluated cases. It is affected by the prevalence of the disease. The diagnostic accuracy generally increases as the prevalence grows. Therefore, diagnostic accuracy or the effectiveness of a test should be considered in a holistic manner by looking at multiple evaluation metrics.

In the discussion, validation in a clinical setting should be clearly stated as the ultimate goal of an ophthalmic AI system is to enhance patient care in a real-world clinical environment. The generalizability of the developed AI system to external validation sets not used for AI development should be evaluated.

Comparisons across reported AI systems are challenging as each system's performance has been tested using different methodologies on different populations, with

different input datasets. In a fair comparison, AI systems need to be subjected to the same independent test set that is representative of the target population, using the same performance metrics. As this is not always possible, the strengths and weaknesses of each paper should be reported.

Current AI systems are capable of diagnosing and staging ocular diseases from images, such as color fundus photographs, OCT, or the visual field. Most algorithms have been tested on limited datasets with relatively homogenous patient populations and a selected set of imaging devices with inclusion criteria that exclude complex pathologies or poor-quality input data for study purposes. Real-world ocular images inevitably include indecipherable scans from diseases not yet medically identified or an overlap of multiple diseases. Future studies should focus on validating algorithms on real-world ocular images from multiple patient populations and using various types and versions of imaging machines.

As the AI field grows in ophthalmology, there is a strong demand for a standardized reporting format and a consensus for judgment criteria. These standards are being established in multiple fields of research and will help clinicians and scientific paper peer reviewers interpret data and apply the results to their research and clinic.

Compliance with ethical standards

Conflict of interest YL, DW, and LP are Google employees and own Alphabet stock, and are inventors on patents for machine learning for medical imaging. DT is a co-inventor of a patent for deep learning system in retinal diseases.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye. 2018;32:1138–44.
2. He J, Cao T, Xu F, Wang S, Tao H, Wu T, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. Eye. 2020;34:572–6.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
4. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318:2211–23.
5. Redd TK, Campbell JP, Brown JM, Kim SJ, Ostmo S, Chan RVP, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br J Ophthalmol. 2018. https://doi.org/10.1136/bjophthalmol-2018-313156.
6. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. 2018;136:803–10.
7. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012;12:8.
8. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: user's guides to the medical literature. JAMA. 2019;322:1806–16.
9. Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus age-related macular degeneration. Ophthalmol Retin. 2017;1:322–7.
10. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial interlligence. JAMA. 2019;322:1765–6.
11. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31.