Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: A systematic review

Amelia Bai\(^1,2,3\), Christopher Carty\(^4,5\), Shuan Dai\(^3,4\)

Abstract:

**PURPOSE:** Artificial intelligence (AI) offers considerable promise for retinopathy of prematurity (ROP) screening and diagnosis. The development of deep-learning algorithms to detect the presence of disease may contribute to sufficient screening, early detection, and timely treatment for this preventable blinding disease. This review aimed to systematically examine the literature in AI algorithms in detecting ROP. Specifically, we focused on the performance of deep-learning algorithms through sensitivity, specificity, and area under the receiver operating curve (AUROC) for both the detection and grade of ROP.

**METHODS:** We searched Medline OVID, PubMed, Web of Science, and Embase for studies published from January 1, 2012, to September 20, 2021. Studies evaluating the diagnostic performance of deep-learning models based on retinal fundus images with expert ophthalmologists’ judgment as reference standard were included. Studies which did not investigate the presence or absence of disease were excluded. Risk of bias was assessed using the QUADAS-2 tool.

**RESULTS:** Twelve studies out of the 175 studies identified were included. Five studies measured the performance of detecting the presence of ROP and seven studies determined the presence of plus disease. The average AUROC out of 11 studies was 0.98. The average sensitivity and specificity for detecting ROP was 95.72% and 98.15%, respectively, and for detecting plus disease was 91.13% and 95.92%, respectively.

**CONCLUSION:** The diagnostic performance of deep-learning algorithms in published studies was high. Few studies presented externally validated results or compared performance to expert human graders. Large scale prospective validation alongside robust study design could improve future studies.

**Keywords:**
Artificial intelligence, deep learning, diagnosis, retinopathy of prematurity, screening

INTRODUCTION

The concept of artificial intelligence (AI) dates back to the 1950s, when Alan Turing first discussed how to build and test intelligent machines in the paper "computing machinery and intelligence."\(^{[1]}\) It wasn’t until 1956, however, at the seminal conference Dartmouth Summer Research Project on AI, did John McCarthy officially coin the term AI. This conference introduced a computer program designed to mimic the problem solving skills of a human, catalyzing the next 20 years of AI research.\(^{[2]}\) Today, AI is incorporated into many applications for day-to-day life, including speech recognition, photo captioning, language translation, robotics, and even self-driving cars.\(^{[3,4]}\) These applications are made possible through the use of deep learning, an advanced form of AI which self-learns from large training sets to program itself to perform certain tasks.\(^{[5]}\) The application of AI has gained popularity in the medical diagnostic field, and promising outcomes have resulted from deep-learning screening algorithms in Ophthalmology.

There has been particular success in AI screening for diabetic retinopathy, with several groups reporting deep-learning algorithms detecting diabetic retinopathy at sensitivities...
and specificities of 83%–90% and 92%–98% respectively.\cite{7,8} Moreover, the successful validation of these algorithms has seen progression to “real-world” implementation of screening programs through prospective evaluation. One such study produced a sensitivity of 83.3% and specificity of 92.5% in detecting referable diabetic retinopathy in a prospective evaluation.\cite{9} Similar promising results are being reported by many other groups utilizing deep learning for the diagnosis of other ophthalmic conditions including diabetic macula edema,\cite{10} age-related macular degeneration,\cite{11} glaucoma,\cite{12} and retinopathy of prematurity (ROP).\cite{13,14}

ROP is a retinal vascular proliferative disease affecting premature infants whose diagnosis is dependent on timely screening. Globally, it is estimated that at least 50,000 children are blind from ROP,\cite{15} and it remains the leading cause of preventable childhood blindness.\cite{16} Advances in retinal imaging means disease is now easily identifiable by retinal photographs, making it a perfect candidate for deep learning. As survival rates of premature infants continue to increase with medical advances,\cite{17} the demand for ROP screening is rapidly exceeding the capacity of available specialist ophthalmologists. For this reason, reports of deep-learning models matching or exceeding human experts in ROP diagnostic performance have generated considerable interest. It remains fundamental, however, that this enthusiasm does not overrule the need for critical appraisal as a missed diagnosis of ROP can result in significant sequelae such as blindness. Therefore, any deep-learning screening algorithm will need to show high diagnostic performance, high sensitivity, be generalizable, and be applicable to the real-world setting. In anticipation of deep-learning diagnostic tools becoming implemented into clinical practice, it is judicious to systematically review the body of evidence supporting AI screening for ROP. This systematic review aims to critically appraise the current state of diagnostic performance of deep-learning algorithms for ROP screening, with particular consideration for study design, algorithm development, type of validation, performance compared to clinicians, and diagnostic accuracy.

**Methods**

**Search strategy and selection criteria**

Studies that developed or validated a deep learning model for the diagnosis of ROP and compared accuracy of algorithm diagnoses to ROP experts were included in this systematic review. We searched MEDLINE-Ovid, PubMed, Web of Science, and Embase for studies published from January 1, 2012 to September 20, 2021. The full search strategy for each database is available in Appendix 1. The cutoff of January 1, 2012 was prespecified based on an important breakthrough made with the development of deep-learning approaches in the model AlexNet.\cite{17} The search was first performed on July 10, 2020, revised on May 23, 2021 and updated on September 20, 2021. Manual searches of bibliographies and citations from included studies were also completed to identify any additional articles potentially missed by searches.

Eligibility assessment was conducted by two reviewers who independently screened titles and abstracts of search results. Only studies aiming to identify through AI algorithms the presence of the disease of interest, ROP, were included. We accepted standard-of-care diagnosis, expert opinion or consensus as adequate reference standards to classify the absence or presence of disease. We excluded studies that did not test for diagnostic performance or investigated accuracy of image segmentation rather than disease classification. Studies which assessed the ability to classify disease severity were accepted if they incorporated primary results of disease detection. Review articles, conference abstracts, and studies that presented duplicate data were excluded. We assessed the risk of bias in patient selection, index test, reference standard, and flow and timing of each study using QUADAS-2.\cite{18} Full assessment of bias can be found in Appendix 2.

This systematic review was completed following the recommendations of the Preferred Reporting Items for Systematic reviews and Meta-Analyses\cite{19} statement and the research question was formulated according to the CHARMS\cite{20} checklist for systematic reviews of prediction models. Methods of analysis and inclusion criteria were specified in advance.

**Data analysis**

Data were extracted independently by two reviewers (AB and SD) using a predefined data extraction sheet, followed by cross-checking. Any discrepancies were discussed with a third reviewer (CC). Demographics and sample size (gestational age [GA], birth weight, number of participants, and number of images), data characteristics (data source, inclusion and exclusion criteria, and image augmentation), algorithm development (architecture, transfer learning, and number of images for training and tuning), algorithm validation (reference standard, number of experts, same method for assessing reference standard, and internal and external validation), and results (sensitivity, specificity, area under the receiver operating characteristic curve for algorithm (AUROC), human graders, and external validation if applicable) were sought. Two papers produced different algorithms from different data sets or with different identification tasks and were therefore recorded as separate algorithms in the results section.\cite{21,22} Data from all 12 papers were included and any missing information was recorded. In the case where sensitivity and specificity were not explicitly recorded but could be calculated from a confusion matrix, the calculated results were included.

**Results**

Our search identified 175 records, of which 99 were screened [Figure 1]. Thirty full-text articles were assessed for eligibility and 12 studies were included in the systematic review.\cite{12,13,21-31} Fifty studies were excluded due to no test of diagnostic performance,\cite{32-39} no classification task,\cite{40-42} no internal validation,\cite{23,43} no AI algorithm,\cite{44} and not based on standard clinical care.\cite{45}
The average sensitivity and specificity were recorded in four of those studies and averaged 95.72% and 98.15%, respectively. One study compared human grader performance to the AI algorithm revealing similar sensitivities (94.1% AI, 93.5% human) and specificities (99.3% AI, 99.5% human) of ROP diagnostic performance. Two of the five studies underwent external validation revealing an average sensitivity and specificity of 60% and 88.3%, respectively, for detecting the presence of disease. The seven other studies determined the ability of their algorithm to detect the presence of plus disease. Among these, six studies measured AUROC, with which the average was 0.98. The sensitivity and specificity for detecting plus disease recorded from six studies were 91.13% and 95.92%, respectively. External validation occurred in two of these studies and produced an average sensitivity of 93.45% and specificity of 87.35%. Performance of AI algorithm at detecting pre-plus disease was measured in two articles producing an average sensitivity of 96.2% and specificity of 95.7%.

Discussion
We found that deep-learning algorithms for ROP screening demonstrated sensitivity and specificity metrics that were comparable to neural network algorithms in diabetic retinopathy. Although this estimate supports the application potential for deep-learning algorithms to be implemented as a real-world diagnostic tool, there are several methodological deficiencies that were common across included studies which need to be considered. These include the quality of reference standard, use of sample size calculations, external validation, definition of presence or absence of disease, and the need for prospective evaluation.
| Study | Source of data | Date range | Open-access data | Missing data | Inclusion criteria | Exclusion criteria | Exclusion of poor-quality imaging/image augmentation |
|-------|----------------|------------|------------------|--------------|-------------------|-------------------|--------------------------------------------------|
| Brown et al., 2018 | Retrospective cohort, data collected at multiple hospitals across North America | July 2011-December 2016 | N | NR | NR | Stage 4-5 ROP | Y/NR |
| Chen et al., 2020 | American Trained Algorithm | | | | | | |
| | Routine ROP screening from 9 North American institutions | | | | | | |
| | Nepalese Trained Algorithm | | | | | | |
| | ROP screening program from 4 urban hospitals in Kathmandu, Nepal (Patan Hospital, Kanti Children’s Hospital, Paropakar Maternity and Women’s Hospital, Tilganga Institute of ophthalmology) | | | | | | |
| Hu et al., 2019 | Chengdu women and Children’s Central Hospital | 2014-2017 | NR | NR | NR | NR | N/YNR |
| Huang et al., 2020 | Neonatal intensive care unit of Chang Gung Memorial Hospital, Linkou, Taiwan | 17 Dec 2013-24 May 2019 | NR | NR | NR | NR | N/Y |
| Mao et al., 2020 | Eye Hospital of Wenzhou Medical University | July 2013 - May 2018 | NR | NR | Only images of the posterior retina | NR | N/YNR |
| Ramachandran et al., 2021 | KidROP Bangalore, India | | | | | | |
| | | | | | | | |
| Tan et al., 2019 | ROP (ART-ROP) image library - from four neonatal ICU in Auckland, New Zealand | 2006-2015 | NR | NR | Poor image quality | Y - Not grossly out of focus, not affected by blur | N/Y |
| Tong et al., 2020 | Images collected from Renmin Hospital of Wuhan University Eye Centre | 1 Feb 2012-1 Oct 2016 | NR | NR | 1. Poor image quality, 2. Imaging artefacts 3. Unfocused scans 4. Presence of other disease phenotypes (e.g., retinal haemorrhage) | Y/Y | |
| Wang et al., 2021 | 4 centres in southern China - JSIEC of Shantou University and Chinese University of Hong Kong, Guangdong Women and Children Hospital in Yuxiu branch (Yuxiu) and Panyu branch (Panyu), and the Sixth Affiliated Hospital of Guangzhou Medical University and Qingyuan People’s Hospital | 1 Sept 2018-24 June 2020 | NR | NR | 1. Nonfundus photos or fundus photos taken by imaging devices other than RetCam 2. Infants with other ocular diseases e.g., congenital cataract, retinoblastoma, or persistent hyperplastic primary vitreous, and 3. any images with disagreeing labels | Y/Y | |
| Wang et al., 2018 | Images captured during routine clinical ROP Screening from Chengdu Women and Children's Central Hospital | Jan 2018 | NR | NR | | | |

Contd...
It is important to differentiate these diagnostic targets as the clinical implication of the findings will differ. In addition, most studies utilized a reference standard graded by on average 2–3 experts with only one study producing a reference standard diagnosed by 5 clinicians per image.\[^{31}\] It is well reported that there is a significant amount of intergrader variability in ROP diagnosis due to its subjective nature;\[^{47,48}\] therefore, caution needs to be taken in recognizing the potential for grader bias in studies utilizing only a few expert graders.

Second, there was a large variety in the number of images used to train each algorithm, ranging from 289\[^{27}\] to 39,029 images.\[^{29}\] Convolutional neural networks learn by computing the error between the machine’s output and the image diagnosis; hence, the more images used to train a machine the smaller the error of its diagnostic output.\[^{6}\] For this reason, the studies that had sample sizes in the ten-thousands were likely to have more reliable results than those that were trained off hundreds or thousands of images. Nonetheless, no studies reported formal sample size calculations to ensure sufficient sizing of studies. Despite the challenge of sample size calculations in the context of AI algorithms, it remains a principal component of any study design and only one paper reported sample size as a limitation in their study.\[^{25}\] Future studies should consider formulating sample size calculations to justify the number of images required for algorithm design.

Thirdly, exclusion of poor-quality images or image augmentation may impact the performance of these deep-learning algorithms in the real-world clinical setting. This is a factor which may limit the diagnostic performance of an algorithm as high quality images correlates to high quality diagnoses and smaller algorithm errors.\[^{6}\] For this reason, it is understandable that most papers will exclude poor quality images; however, it is important to keep this within reason. Quality of images used to train an algorithm should correspond to the quality of images taken in the clinical setting so that algorithm performance may equate to its real-life performance. It is also for this reason that external validation of an algorithm using an image set outside of the training image set is crucial to determine the generalizability of a study. Only five of the twelve studies completed external validation of which all but one study, showing equivalent performance, revealed inferior algorithm performance compared to their test set. This finding highlights the need for an out-of-sample external validation in these screening algorithms to better understand how the algorithm will perform in the clinical setting.

Fourth, the ground truth or reference standard labels were mostly derived from data collected for other purposes such as a database of ROP images or retrospective routine clinical care notes. Although there exists an internationally accepted guideline for defining presence and stage of ROP, the International Classification of Retinopathy of Prematurity revisited (ICROP)\[^{49}\] (more recently updated in a 2021 version\[^{50}\]), only five studies specifically mentioned the ICROP in their methods for defining the reference standard. As ICROP
### Table 2: Patient demographics and sample size for the 12 included studies

| Study                        | Participants | Sample size | Number of Images with ROP | Number of Images with plus disease | Number of Images with preplus |
|------------------------------|--------------|-------------|---------------------------|-----------------------------------|------------------------------|
| Brown et al., 2018           | NR           | NR          | 898                       | 5511                              | N/A                          |
| Chen et al., 2020            | American Trained Algorithm: 26.6 (2.2; NR) | 856.2 (293.7; NR) | NR                        | 5943                              | NR                           |
| Nepalese Trained Algorithm   | 32.6 (2.8; NR) | 1949.6 (495.8; NR) | NR                        | 5049                              | NR                           |
| Hu et al., 2019              | 32 (NR; NR)  | 1964 (NR; NR) | 720                       | 2668                              | 1184                         |
| Huang et al., 2020           | 27.3 (1.8; NR) | 936.4 (229.8; NR) | NR                        | 10235                             | NR                           |
| Mao et al., 2020             | 31.3 (2.1; NR) | NR          | 3021                      | 6161                              | N/A                          |
| American Trained Algorithm   | 1643 (419.5; NR) | NR          | 557 (stage 1 ROP)         | 722 (stage 2 ROP)                 | N/A                          |
| Ramachandran et al., 2021    | 32.4 (1.1; NR) | 1350 (240; NR) | NR                        | 150                               | 289                          |
| Tan et al., 2019             | 1280 (226; NR) | NR          | 4926                      | N/A                               | 1638                         |
| Tong et al., 2020            | NR           | NR          | 3487 suitable for training | N/A                               | 3006                         |
| Wang et al., 2021            | 32.9 (3.1; NR) | 1925 (774; NR) | NR                        | 8652                              | N/A                          |
| Wang et al., 2018            | Id - Net     | NR*         | NR                        | 52249                             | NR                           |
|                             | Gr - Net     | NR          | NR                        | 1273 total for both Id - Net and Gr - Net                        | N/A                          |
|                             |              | NR          | 605 for developing data   | 6917                              | N/A                          |
|                             |              | NR          | 264 for data for expert training | 5967 (developing)             | N/A                          |
|                             |              | NR          | 404 data from web         | 293 (expert comparison)           | N/A                          |
|                             |              | NR          | 2361 (expert comparison)  | 657 (data from web)              | N/A                          |
|                             |              | NR          | 4908 (from web)           | 404 data from web                | N/A                          |
|                             |              | NR          | 5089                      | 5139 (developing)               | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 293 (expert comparison)          | N/A                          |
|                             |              | NR          | 4139 (developing)         | 657 (data from web)             | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 404 data from web                | N/A                          |
|                             |              | NR          | 5089                      | 5139 (developing)               | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 657 (data from web)             | N/A                          |
|                             |              | NR          | 4139 (developing)         | 404 data from web                | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 5139 (developing)               | N/A                          |
|                             |              | NR          | 4139 (developing)         | 657 (data from web)             | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 404 data from web                | N/A                          |
|                             |              | NR          | 5089                      | 5139 (developing)               | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 657 (data from web)             | N/A                          |
|                             |              | NR          | 4139 (developing)         | 404 data from web                | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 5139 (developing)               | N/A                          |
|                             |              | NR          | 4139 (developing)         | 657 (data from web)             | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 404 data from web                | N/A                          |
|                             |              | NR          | 5089                      | 5139 (developing)               | N/A                          |
|                             |              | NR          | 293 (expert comparison)   | 657 (data from web)             | N/A                          |
|                             |              | NR          | 4139 (developing)         | 404 data from web                | N/A                          |

*Data represented as bar graph distribution, unable to calculate mean. NR: Not recorded, N/A: Not applicable, SD: Standard deviation, BW: Birth weight, GA: Gestational age, SD: Standard deviation, ROP: Retinopathy of prematurity
Table 3: Details of algorithm development for the 12 included studies

| Study               | Algorithm name | Algorithm architecture                          | Transfer learning applied | Number of images for training/tuning |
|---------------------|----------------|-----------------------------------------------|---------------------------|--------------------------------------|
| Brown et al., 2018  | NR             | U-Net architecture                           | Yes                       | 4409/1102                            |
| Chen et al., 2020   | American Trained Algorithm | NR | ImageNet and Pytorch | Yes (from ResNet) | 5255/NR                     |
|                     | Nepal Trained Algorithm | NR | ImageNet and Pytorch | Yes (from ResNet) | 4802/NR                     |
| Hu et al., 2019     | NR             | ImageNet and TensorFlow (Inception-v2, VGG-16, ResNet-50) | Yes (from ImageNet) | 2068/300                            |
| Huang et al., 2020  | NR             | Tensorflow                                   | No                        | 10,235/1137                         |
| Mao et al., 2020    | NR             | U-Net and DenseNet                           | Yes (from ImageNet)       | 5711/NR                              |
| Ramachandran et al., 2021 | NR | Darknet-53                                   | Yes (from ImageNet)       | 289/32 (then retrained by 96 images for final model) |
| Tan et al., 2019    | ROP.AI         | TensorFlow’s Inception-v3                    | Yes                       | 80% of 6974/NR                     |
| Tong et al., 2020   | NR             | Faster R-CNN+TensorFlow                      | Yes (from ResNet)         | 90% of 26,459/10% of 26,459        |
| Wang et al., 2021   | J-PROP         | NR*                                          | Yes (from Res-Unet)       | 75% (39,029)/10% (5140)            |
| Wang et al., 2018   | DeepROP        | Tensorflow - Inception-BN Network            | Yes (from ImageNet)       | 17665/NR                            |
| Yildiz et al., 2020 | I-ROP ASSIST   | NR*                                          | Yes (from U-Net)          | 5512/NR                             |
| Zhang et al., 2018  | CAD-R          | NR*                                          | Yes (from VGG-16)         | 17801/NR                            |

*Specific architecture not recorded, however CNN used. NR: Not recorded, AI: Artificial intelligence, ROP: Retinopathy of prematurity, CAD-R: Computer-aided diagnosis system for ROP, R-CNN: Region-convolutional neural networks, I-ROP: Indian ROP, BN: Batch normalised, VGG: Visual geometry group

acts as the universally adopted diagnostic criteria for grading ROP, it is safe to assume that the other seven studies also used these guidelines; however, the criteria for the presence or absence of disease should always be clearly defined in AI studies.

Finally, only one study completed prospective evaluation of their algorithm, a process that is vital to assess the performance on real-world implications. The majority of studies assessed deep learning diagnostic accuracy in isolation, without external validation as mentioned earlier or comparison to experts. Only three studies provided a comparison of AI performance with human performance, allowing for evaluation of real-world application. Without comparison of AI to human performance, the results from the other seven studies are limited in their ability to be extrapolated into health-care delivery. In order for a deep learning diagnostic tool to be applicable in clinical bedside screening, it must perform better or comparable to the gold standard, in this case expert diagnosis. More work is required to validate the performance of AI algorithms in comparison to human graders, ideally using the same external test dataset.

It is clear from this systematic review that there still lacks a well-designed randomized head-to-head comparison of an effective externally validated AI algorithm to human performance in real-time. A study of this magnitude could reveal the possible clinical implications for an algorithm implemented in the clinical setting. For this reason, prospective evaluations of these deep-learning diagnostic tests are crucial to unveil the bounding potential of AI in both diagnostic and therapeutic medicine. We recognize that there is a large “black box” issue in deep learning, where image features learned by an algorithm is unknown to the user. It is for this reason that many clinicians are sceptical to entrust clinical care to AI, especially when the clinical features clinicians are familiar with may not be the same features used by an algorithm. This further emphasizes the need for well executed studies that minimize bias and are thoroughly and transparently reported. Most of the concerns we have highlighted in this review are avoidable with robust design and it remains critical that these AI diagnostic tests are evaluated in the context of its intended clinical pathway.

**Conclusion**

AI has been heralded as a revolutionary technology for many industries, and certainly deep learning algorithms for diagnosis of ROP are no exception. Despite the issues we have highlighted in this systematic review, the performance of the twelve deep-learning algorithms evaluated has been extremely high, with all studies delivering a recordable AUROC above or equivalent to 0.94. These results outline the ability for AI algorithms to perform comparable to or exceeding human experts and provide the groundwork for future large-scale prospective studies. Although there are clear screening and treatment guidelines, ROP disease burden continues to rise as increased survival of preterm infants coincides with advancements in medical care. The inadequate accessibility and number of experienced ophthalmologists continues to limit ROP screening and diagnosis. Consequently, the burden of ROP visual impairment is expected to increase unless a novel strategy such as deep-learning diagnostic algorithms becomes available. There is no doubt that the successful application of AI in ROP will revolutionize disease diagnosis through its high predictive performance and streamlined efficiency. The clinical implications of this implementation into real-world clinical practice is immeasurable, with translation into high accessibility, high quality, timely screening and the
Table 4: Method of algorithm validation for the 12 included studies

| Study                  | Reference standard                                      | If compared to experts, how many? | Same method for assessing reference standard across samples | Type of internal validation                  | Number of images for internal validation | External validation | Number of images for external validation |
|------------------------|---------------------------------------------------------|----------------------------------|-----------------------------------------------------------|---------------------------------------------|------------------------------------------|---------------------|------------------------------------------|
| Brown et al., 2018     | Expert consensus                                        | 3                                | Yes                                                      | Random split sample validation             | 20%                                      | Yes                 | 100                                      |
| Chen et al., 2020      | American Trained Algorithm                              | 3                                | Yes                                                      | 5-fold cross-validation                    | 10% test set                            | Yes                 | 247 images from Nepal                   |
|                        | Expert consensus                                        | 1                                | Yes                                                      | 5-fold cross-validation                    | 10% test set                            | Yes                 | 708 images from America                 |
| Hu et al., 2019        | Expert consensus                                        | 3                                | Yes                                                      | Random split sample validation             | 300                                      | No                  | N/A                                      |
| Huang et al., 2020     | Expert consensus                                        | 3                                | Yes                                                      | 5-fold cross-validation                    | 244                                      | No                  | N/A                                      |
| Mao et al., 2020       | Clinical diagnosis by one ophthalmologist               | 1                                | NR                                                       | Random split                              | 450                                      | No                  | N/A                                      |
| Ramachandran et al., 2021 | Expert consensus                                      | 3                                | Yes                                                      | 80:20 split                               | 161 (67 ROP)                            | No                  | N/A                                      |
| Tan et al., 2019       | Expert ophthalmologist from New Zealand; External images graded by expert ophthalmologist from Hong Kong | 1                                | No - 2 different experts between internal and external validation | 80:20 random split validation             | 20% of 6974                              | Yes                 | 90 (33 plus, 57 normal) + additional 26 preplus images for assessing preplus |
| Tong et al., 2020      | Expert grading (11 retinal experts for first-round screening, 2 senior experts confirmed or corrected labels) | 2                                | No - 11 different first round graders                    | 10-fold cross-validation                   | 9772                                     | No*                 | N/A                                      |
| Wang et al., 2021      | Expert grading (2 junior ophthalmologists labelled, any disagreement submitted to 1 senior ophthalmologist) | 3                                | No - dependent on agreement                              | Random split 75:10:15 (training, validation, test) - but based on a patient-based split policy (i.e., all images of a patient were allocated into the same sub-data set) | 8080                                     | No                  | N/A                                      |
| Wang et al., 2018      | Expert consensus (images included if 2 out of 3 graders agreed, disagreements sent to fourth ophthalmologist) | 3-4                              | Yes                                                      | Random split                              | 298 (for Id-Net), 104 (for Gr-Net)      | Yes - prospective evaluation          | 2361 (total, Id and Gr net)              |
| Yildiz et al., 2020    | Expert consensus                                        | 3                                | Yes                                                      | 5-fold cross-validation                    | 5000                                     | Yes                 | 100 (15 plus, 34 preplus)               |
| Zhang et al., 2018     | Cross validation by one senior ophthalmologist          | 5 (2 senior experts, 2 attending physicians, 1 resident) | Yes                                                      | Random selection                          | 1742 (155 ROP, 1587 without ROP)        | No                  | N/A                                      |

*No external validation, however did measure algorithm versus human graders on another 1227 images collected during routine clinical care. ROP: Retinopathy of prematurity, N/A: Not applicable, NR: Not recorded
Table 5: Summary of results from the 12 included studies

| Study                     | Algorithm Performance | Sens % | Spec % | Area under the ROC curve (AUROC) |
|---------------------------|-----------------------|--------|--------|----------------------------------|
| Detecting Disease         |                       |        |        |                                  |
| Hu et al. 2019            |                       | 96     | 98     | 0.9922                           |
| Zhang et al. 2018         |                       | 94.1   | 99.3   | 0.998                           |
| Chen et al. 2020          |                       |        |        |                                  |
| American Trained Algorithm|                       | NR     | NR     | 0.99                             |
| Nepal Trained Algorithm   |                       | NR     | NR     | 0.96                             |
| Combined (American & Nepal) Trained Algorithm| | NR | NR | 0.99 |

| Study                     | Sens grading % | Spec grading % | AUROC grading |
|---------------------------|----------------|----------------|---------------|
| Detecting Disease & Stage | 96.14±0.87     | 95.95±0.48     | 0.96          |
| Huang et al. 2020         | 91.82±2.03 (stage 1) | 94.5±0.71 (stage 1) | 0.93 (stage 1) |
| Wang et al. 2018          | N/A            | N/A            | N/A           |
| Id-Net                    | 96.64          | 99.33          | 0.995         |
| Gr-Net                    | N/A            | N/A            | N/A           |

| Study                     | Sens Pre-Plus % | Spec Pre-Plus % | AUROC Pre-plus |
|---------------------------|----------------|----------------|----------------|
| Detecting Plus Disease    | 93             | 94             | NR             |
| Brown et al. 2018         | 95.1           | 97.8           | 0.99           |
| Mao et al. 2020           | 99             | 98             | 0.9947         |
| Ramachandran et al. 2021  | N/A            | N/A            | N/A            |
| Tan et al. 2019           | 96.6           | 98             | 0.993          |
| Yildiz et al. 2020        | NR             | NR             | 0.94           |

| Study                     | Sens Pre-Plus % | Spec Pre-Plus % | AUROC Pre-plus |
|---------------------------|----------------|----------------|----------------|
| Detecting Plus & Severity | 91.8           | 97             | 0.983          |
| Wang et al. 2021          | 98.2 (stage)   |                 | 0.998 (stage)  |
| Tong et al. 2020          | 71.3           | 90.7           | NR             |
|                           | 77.8 (“normal” “mild”) | 93.2 (“normal” “mild”) | NR             |
|                           | “semi-urgent” “urgent” | “semi-urgent” “urgent” |               |

Contd...
### Table 5: Contd...

| Study | Human Performance | | | | | External Validation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sens % | Spec % | Sens grading % | Spec grading % | AUROC | Sens % | Spec % | AUROC | | | | |
| Detecting Disease | | | | | | | | | | | | |
| Hu et al. 2019 | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | | | |
| Zhang et al. 2018 | 93.5 | 99.5 | n/a | n/a | n/a | 52 | 99 | 0.96 | | | | |
| Chen et al. 2020 | | | | | | | | | | | | |
| American Trained Algorithm | n/a | n/a | n/a | n/a | n/a | 98/82 (against American/Nepal set) | 96/99 (against American/Nepal set) | 0.99/0.98 (against American/Nepal set) | | | | |
| Combined (American & Nepal) Trained Algorithm | n/a | n/a | n/a | n/a | n/a | | | | | | | |
| Detecting Disease & Stage | | | | | | | | | | | | |
| Huang et al. 2020 | NR | NR | NR | NR | NR | n/a | n/a | n/a | n/a | n/a | n/a | |
| Wang et al. 2018 | NR | NR | NR | NR | NR | | | | | | | |
| Id-Net | NR | NR | NR | NR | NR | 84.91 | 96.9 | NR | n/a | n/a | n/a | |
| Gr-Net | NR | NR | NR | NR | NR | n/a | n/a | n/a | 93.33 (minor vs. severe) | 73.63 (minor vs. severe) | NR | |
| Detecting Plus Disease | | | | | | | | | | | | |
| Brown et al. 2018 | n/a | n/a | n/a | n/a | n/a | 93 | 94 | NR | 100 | 94 | NR | |
| Mao et al. 2020 | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | |
| Ramachandran et al. 2021 | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a | |
| Tan et al. 2019 | n/a | n/a | n/a | n/a | n/a | 93.9 | 80.7 | NR | 81.4 | 80.7 | 0.977 | |
| Yildiz et al. 2020 | n/a | n/a | n/a | n/a | n/a | NR | NR | 0.99 | NR | NR | 0.97 | |
| Detecting Plus & Severity | | | | | | | | | | | | |
| Wang et al. 2021 | 100 (compared to J-PROP on same dataset: 98.4) | 99.8 (compared to J-PROP on same dataset: 98.4) | 91.7 (stage) (compared to J-PROP on same dataset: 97.9) | 99.1 (stage) (compared to J-PROP on same dataset: 97.4) | NR | n/a | n/a | n/a | n/a | n/a | n/a | |
| Tong et al. 2020 | 74.8 (expert 1), 65.9 (expert 2) (for grading “normal” “mild” “semi-urgent” “urgent”) | 93.4 (expert 1), 92.3 (expert 2) (for grading “normal” “mild” “semi-urgent” “urgent”) | n/a | n/a | n/a | NR | n/a | n/a | n/a | n/a | n/a | |
significant reduction in cost of screening. AI will therefore become ubiquitous and indispensable for ROP screening, and it is important that high quality research continues to aid the translation of this transformative technology in order to reduce the incidence of visual loss and blindness from this preventable disease.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES
1. Turing A. Computing machinery and intelligence. Mind 1950;49:433-60.
2. McCarthy J, Minsky, M, Rochester N, Shannon C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence 1955. AI Magazine 2006;27:12.
3. Wu J, Yilmaz E, Zhang M, Li H, Tan KC. Deep spiking neural networks for large vocabulary automatic speech recognition. Front Neurosci 2020;14:199.
4. Popel M, Tomkova M, Tomek J, Kaiser L, Uszkoreit J, Bojar O, et al. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. Nat Commun 2020;11:4381.
5. Fayyad J, Jaradat MA, Gruyer D, Najhara H. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. Sensors (Basel) 2020;20:42220.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
7. 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.
8. Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z, et al. Artificial intelligence-enabled screening for diabetic retinopathy. A real-world, multicenter and prospective study. BMJ Open Diabetes Res Care 2020;8:e001596.
9. De Fauw J, Ledsam JR, Romero-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.
10. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol 2017;135:1170-6.
11. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 2018;125:1199-206.
12. 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.
13. Tan Z, Simkin S, Lai C, Dai S. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. Transl Vis Sci Technol 2019;8:23.
14. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr Res 2013;74 Suppl 1:35-40.
15. Gilbert C. Retinopathy of prematurity: A global perspective of the epidemics, population of babies at risk and implications for control. Early Hum Dev 2008;84:77-82.
16. Valentine PH, Jackson JC, Kalina RE, Woodrum DE. Increased survival of low birth weight infants: Impact on the incidence of retinopathy of prematurity. Pediatrics 1989;84:442-5.
17. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
18. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Int J Surg 2021;88:105906.
20. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. PLoS Med 2014;11:e1001744.
21. Chen JS, Coyer AS, Ostmo S, Sonmez K, Bajina Maya S, Pradhan E, et al. Deep learning for the diagnosis of stage in retinopathy of prematurity: Accuracy and generalizability across populations and cameras. Ophthalmol Retina 2021;5:1027-35.
22. Wang J, Ju R, Chen Y, Zhang L, Hu J, Wu Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBioMedicine 2018;35:361-8.
23. Campbell JP, Singh P, Redd TK, Brown JM, Shah PK, Subramanian P, et al. Applications of artificial intelligence for retinopathy of prematurity screening. Pediatrics 2021;147:e2020016618.
24. Hu J, Chen Y, Zhong J, Ju R, Yi Z. Automated analysis for retinopathy of prematurity by deep neural networks. IEEE Trans Med Imaging 2019;38:269-79.
25. Huang YP, Absanta H, Kang EY, Chen KJ, Hung WS, Lai CC, et al. Automated detection of early-stage ROP using a deep convolutional neural network. Br J Ophthalmol 2021;105:1099-103.
26. Mao J, Luo Y, Liu L, Lao J, Shao Y, Zhang M, et al. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. Acta Ophthalmol 2020;98:e339-45.
27. Ramachandran S, Niyas P, Vinekar A, John R. A deep learning framework for the detection of Plus disease in retinal fundus images of preterm infants. Bioeyben Biomed Eng 2021;41:362-75.
28. Tong Y, Lu W, Deng QQ, Chen C, Shen Y. Automated identification of retinopathy of prematurity by image-based deep learning. Eye Vis (Lond) 2020;7:40.
29. Wang J, Ji J, Zhang M, Lin JW, Zhang G, Gong W, et al. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. JAMA Netw Open 2021;4:e218758.
30. Yildiz VM, Tian P, Yildiz I, Brown JM, Kalpathy-Cramer J, Dy J, et al. Plus disease in retinopathy of prematurity: Convolutional neural network performance using a combined neural network and feature extraction approach. Transl Vis Sci Technol 2020;9:10.
31. Zhang Y, Wang L, Wu Z, Zeng J, Chen Y, Tian R, et al. Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. Ieee Access 2019;7:10232-41.
32. Coyer AS, Swan R, Campbell JP, Ostmo S, Brown JM, Kalpathy-Cramer J, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. Ophthalmol Retina 2019;3:444-50.
33. Greenland MF, Danford ID, Shahrawat M, Ostmo S, Brown J, Kalpathy-Cramer J, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. J AAPOS 2020;24:160-2.
34. Gupta K, Campbell JP, Taylor S, Brown JM, Ostmo S, Chan RVP, et al. A Quantitative severity scale for retinopathy of prematurity using deep learning to monitor disease regression after treatment. JAMA Ophthalmol 2019;137:1029-36.
35. Redd T, Campbell J, Brown J, Kim S, Ostmo S, Chan R, et al. Utilization of a deep learning image assessment tool for epidemiologic surveillance of retinopathy of prematurity. Invest Ophthalmol Vis Sci 2019;60:580-4.
36. Smith K, Kim S, Goldstein I, Ostmo S, Chan R, Brown J, et al. Quantitative analysis of aggressive posterior retinopathy of prematurity using deep learning. Invest Ophthalmol Vis Sci 2019;60:4759.
37. Taylor S, Kishan G, Campbell P, Brown J, Ostmo S, Chan R, et al. Invest Ophthalmol Vis Sci 2018; 59: 3937.
38. Wallace DK, Zhao Z, Freedman SF. A pilot study using “ROPool” to quantify plus disease in retinopathy of prematurity. J AAPOS 2019;23:28-30.
Bai, et al.: Performance of deep-learning AI algorithms in detecting ROP

2007;11:381-7.

39. Wang J, Zhang G, Lin J, Ji J, Qiu K, Zhang M. Application of standardized manual labeling on identification of retinopathy of prematurity images in deep learning. Zhonghua Shiyan Yanke Zazhi 2019;37:653-7.

40. Campbell J, Chan R, Ostmo S, Anderson J, Singh P, Kalpathy-Cramer J, Chiang M. Analysis of the relationship between retinopathy of prematurity zone, stage, extent and a deep learning-based vascular severity scale. Invest Ophthalmol Vis Sci 2020;61:2193.

41. Choi RY, Brown JI, Kalpathy-Cramer J, Chan RV, Ostmo S, Chiang MF, et al. Variability in plus disease identified using a deep learning-based retinopathy of prematurity severity scale. Ophthalmol Retina 2020;4:1016-21.

42. Ramachandran S, Kochitty S, Vinekar V, John R. A fully convolutional neural network approach for the localization of optic disc in retinopathy of prematurity diagnosis. J Intell Fuzzy Syst 2020;38:6269-78.

43. Worrall DE, Wilson C, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks. In: Deep Learning and Data Labeling for Medical Applications. DLMIA, LABELS; 2016. p. 68-76.

44. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, Bozkurt A, Erdogmus D, Kalpathy-Cramer J, et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the “i-ROP” system and image features associated with expert diagnosis. Transl Vis Sci Technol 2015;4.5.

45. Touch P, Wu Y, Kihara Y, Zepeda E, Gillette T, Cabrera M, et al. Development of AI deep learning algorithms for the quantification of retinopathy of prematurity. J Invest Med 2019;67:209.

46. Wang S, Zhang Y, Lei S, Zhu H, Li J, Wang Q, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: A systematic review and meta-analysis of diagnostic test accuracy. Eur J Endocrinol 2020;183:41-9.

47. Gschlößler A, Stifter E, Neumayer T, Moser E, Papp A, Pircher N, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. Am J Ophthalmol 2015;160:553-60.e3.

48. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, Bozkurt A, Erdogmus D, Kalpathy-Cramer J, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. JAMA Ophthalmol 2016;134:651-7.

49. International Committee for the Classification of Retinopathy of Prematurity. The International classification of retinopathy of prematurity revisited. Arch Ophthalmol 2005;123:991-9.

50. Chiang MF, Quinn GE, Fielder AR, Ostmo SR, Paul Chan RV, Berrocal A, et al. International classification of retinopathy of prematurity, third edition. Ophthalmology 2021;128:c51-68.
Appendix 1: Full search strategy

We show the search strategy for:

- a. Medline OVID
- b. Pubmed
- c. Web of science
- d. Embase

**Medline OVID**
- “Retinopathy of prematurity” or “ROP” and
- “Diagnosis” or “screening” and
- “Artificial intelligence” or “deep learning” or “convolutional neural networks”

**PubMed**
- “Retinopathy of prematurity” or “ROP” and “diagnosis” or “screening” AND “artificial intelligence” or “deep learning” or “convolutional neural network”

**Web of science**
- TI = (diagnosis or screening or classification) and
- TS = (artificial intelligence or machine learning or deep learning or convolutional neural network) and
- TI = (retinopathy of prematurity or ROP)

**Embase**
- “Retinopathy of Prematurity” or “ROP” or “plus disease” and
- “Diagnosis” or “screening” or “classification” and
- “Artificial intelligence” or “deep learning” or “convolutional neural network” or “machine learning”

ROP: Retinopathy of prematurity

Appendix 2: Methodological quality assessment of bias for included studies using QUADAS-2[18]

| Study            | Domain 1A | Domain 1B | Domain 2A | Domain 2B | Domain 3A | Domain 3B | Domain 4A |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Brown et al. 2018| Unclear   | Low       | Low       | Low       | Low       | Low       | Low       |
| Chen et al. 2020 | Low       | Low       | Low       | Low       | Unclear   | Low       | Low       |
| Hu et al. 2019   | Unclear   | Unclear   | Low       | Low       | Low       | Low       | Low       |
| Huang et al. 2020| Low       | Low       | Low       | Low       | Low       | Low       | Low       |
| Mao et al. 2020  | High      | Unclear   | Low       | Low       | High      | Low       | Unclear   |
| Ramachandran et al. 2021 | High | Low       | Low       | Low       | High      | Low       | Low       |
| Tan et al. 2019  | Low       | Low       | Low       | Low       | Low       | Low       | Low       |
| Tong et al. 2020 | Low       | Low       | Low       | Low       | High      | High      | High      |
| Wang et al. 2021 | Low       | Low       | Low       | Low       | Unclear   | Low       | Unclear   |
| Wang et al. 2018 | Low       | Low       | Low       | Low       | Low       | Low       | Low       |
| Yildiz et al. 2020| High      | Low       | High      | Low       | Low       | Low       | Low       |
| Zhang et al. 2018| High      | Low       | Low       | Low       | Low       | Low       | Unclear   |