ORIGINAL ARTICLE

# AI FOR LITTLE EYES: A SYSTEMATIC REVIEW OF DEEP LEARNING IN EVALUATING PEDIATRIC CATARACT

**Tjoeng E[1], Annisa ND[2]**
[1]Faculty of Medicine Universitas Indonesia, Jakarta, Indonesia
[2]JEC Eye Hospitals and Clinics, Jakarta, Indonesia
*Email: eric_1995@msn.com*

## ABSTRACT

***Background***: *Pediatric or congenital cataract (CC) is a leading cause of visual impairment and blindness in children worldwide. Deep learning (DL), a subfield of artificial intelligence, has the potential to enhance diagnosis, treatment, and outcomes in various medical fields.*

***Research Objectives***: *summarize and evaluate the diagnostic and prediction capabilities of DL algorithms for CC.*

***Methods***: *From 1st February to 25th March 2023, a literature search was conducted in databases such as PubMed, ScienceDirect, EMBASE, and EBSCO, as well as alternative sources such as Google Scholar. Search terms included "pediatric/congenital cataract", "artificial intelligence", "deep learning", "convolutional neural network", "diagnosis", "screening", "prediction" and other relevant synonyms. Quality assessment of studies were assessed based on CONSORT-AI and QUADAS-2. Outcomes extracted included accuracy, sensitivity, specificity, and area under the curve (AUC).*

***Results***: *Out of 69 studies screened, five studies with different study designs, dataset sizes, and type of DL algorithms employed were included in the systematic review. Most studies employed DL to analyze slit-lamp images to diagnose CC, while one study utilized DL to predict existence of CC from several risk factors. In silico, most studies demonstrated high accuracy and validity of DL algorithms in detecting and predicting CC; however, DL algorithm is not as accurate in diagnosing CC when compared to human counterparts. These studies had limited generalizability given the homogenous population.*

***Conclusion***: *DL shows potential as an adjunct tool for ophthalmologists to improve diagnosis and, therefore, treatment decisions for CC, particularly in remote and underdeveloped regions with limited medical resources.*

***Keywords***: *pediatric cataract, deep learning, diagnosis, prediction*

## INTRODUCTION

According to the World Health Organization (WHO)[1], pediatric cataract or congenital cataract (CC) is a leading cause of childhood blindness, affecting approximately 200,000 children worldwide with a prevalence of 4.24 per 10,000 live births.[2,3] Although cataract surgery is a safe and effective treatment, visual outcomes in children with cataract can be suboptimal due to factors such as amblyopia, secondary glaucoma, and posterior capsule opacification.[4] However, detecting CC at an early stage is challenging since the progression is often asymptomatic and difficult for parents to identify.[5] Moreover, in some settings, access to ophthalmic care may be limited, resulting in delayed diagnosis and treatment.[6,7] Therefore, accurate early-stage diagnosis is crucial to enable ophthalmologists to arrange appropriate and timely treatment, optimizing outcomes and minimizing the risk of complications.

In recent years, artificial intelligence (AI) has found application in the field of ocular disease diagnosis. A subset of AI, deep learning (DL) through convolutional neural network (CNN) has emerged as an exceptionally powerful tool for medical image analysis and classification. CNNs, inspired by the visual cortex of cats[8], could extract relevant high-level features directly from raw images without the need for extensive pre-processing or expert knowledge. In ophthalmology, DL has been used for tasks such as retinal disease detection, glaucoma diagnosis and anterior segment analysis.[9–11] Aside from CNN, machine learning has also been used in creating predictive models based on structured data including demographics and clinical features[12,13] which could prove to be useful in regions with insufficient access to medical resources for disease screening.

This systematic literature review aims to comprehensively assess the existing evidence and identify knowledge gaps pertaining to the utilization of DL techniques in the diagnosis and prediction of CC.

## METHODS

A comprehensive search was conducted to identify relevant studies on the use of deep learning in diagnosing and predicting pediatric cataract. The search was performed from 1st February to 25th March 2023. The following databases were searched: PubMed, ScienceDirect, EMBASE, and EBSCO. Additionally, alternative sources such as Google Scholar were also searched to ensure a thorough coverage of the literature. Search terms included “pediatric/congenital cataract”, “artificial intelligence", "deep learning", "convolutional neural network", “diagnosis”, "screening", "prediction" and other relevant synonyms (Table 1). The search was also limited to articles published in English with full text availability. No limit was set on the year of publication. Table 1 summarized the search strategy employed in this paper.

The search results were then imported into reference management software, EndNote 20 (Clarivate, Philadelphia), to facilitate the study selection process. Duplicate articles were identified and removed. Two independent reviewers screened the titles and abstracts of the remaining articles to assess their eligibility based on predetermined inclusion and exclusion criteria. The full texts of potentially eligible articles were retrieved and further assessed for inclusion. Any discrepancies between the two reviewers were resolved through internal discussion.

**Table 1. Search terms for each database**

| Databases | Search Terms |
|---|---|
| **PubMed** | (("artificial intelligence" or "deep learning" or "convolutional neural network") AND ("diagnosis" or "screening" or "prediction")) AND (("congenital cataract"[All Fields]) OR ("pediatric cataract"[All Fields])). Filters: Free full text, English |
| **Embase** | ('artificial intelligence'/exp OR 'artificial intelligence' OR 'deep learning'/exp OR 'deep learning' OR 'convolutional neural network'/exp OR 'convolutional neural network') AND ('diagnosis'/exp OR 'diagnosis' OR 'screening'/exp OR 'screening' OR 'prediction'/exp OR 'prediction') AND ('congenital cataract'/exp OR 'congenital cataract' OR 'pediatric cataract' OR 'paediatric cataract') |
| **ScienceDirect** | (("artificial intelligence" OR "deep learning" OR "convolutional neural network") AND ("diagnosis" OR "screening" OR "prediction") AND ("congenital cataract" OR ("pediatric cataract" OR "paediatric cataract"))) |
| **EBSCO** | ((diagnosis or diagnosing or diagnostics) OR (screening or early detection) OR (prediction)) AND (congenital cataract OR pediatric cataract) AND ((artificial intelligence or ai or a.i) OR (deep learning or machine learning or artificial neural network) OR convolutional neural network) |

Studies were included if they met the following criteria: (1) studies focused on the application of DL algorithms in the context of screening, diagnosing, and predicting pediatric cataract, (2) the studies had to have details on the datasets used, diagnosis, prediction and/or grading criteria of pediatric cataract along with the number of research objects (such as images or cases) in each group, (3) the studies should describe the DL algorithms used for diagnosing, grading and predicting pediatric cataract and report evaluation metrics such as accuracy, sensitivity, specificity and area under the receiver operating characteristics curve, AUC or AUROC. Studies were excluded if they were review articles, conference abstracts, or editorials.

Data extraction form was formulated based on the MINimum Information for Medical AI Reporting (MINIMAR)[14]. Two assessors then independently collected relevant information from the chosen articles. Various information including the author, publication year, population included, study setting, data source, definition and grade of pediatric cataract, dataset characteristics, DL algorithm, training, validation and test datasets, and all diagnostic values in training, validation, and testing datasets (accuracy, sensitivity, specificity, and AUC) were collected. Any discrepancies in data extraction were resolved through discussion and consensus. The findings from the included studies were synthesized narratively. A descriptive summary of the characteristics and main findings of each study was provided.

To assess for risk of bias in the selected diagnostic studies, the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) framework[15] was utilized. While, for clinical trials, CONSORT-AI[16] were used. This assessment was performed by two independent reviewers, and any disagreements were resolved through discussion.

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[17] to ensure transparent and accurate reporting of the search strategy, study selection process, data extraction, and quality assessment.

## RESULTS

The search strategy (Figure 1) identified a total of 101 studies from various databases and websites, of which 70 were screened. Sixty-three studies were excluded as they were review articles (n= 7), conferences abstracts (n= 1), book chapters (n= 9) and studies that did not focus on CC (n= 46). A total of five articles were included in this study.

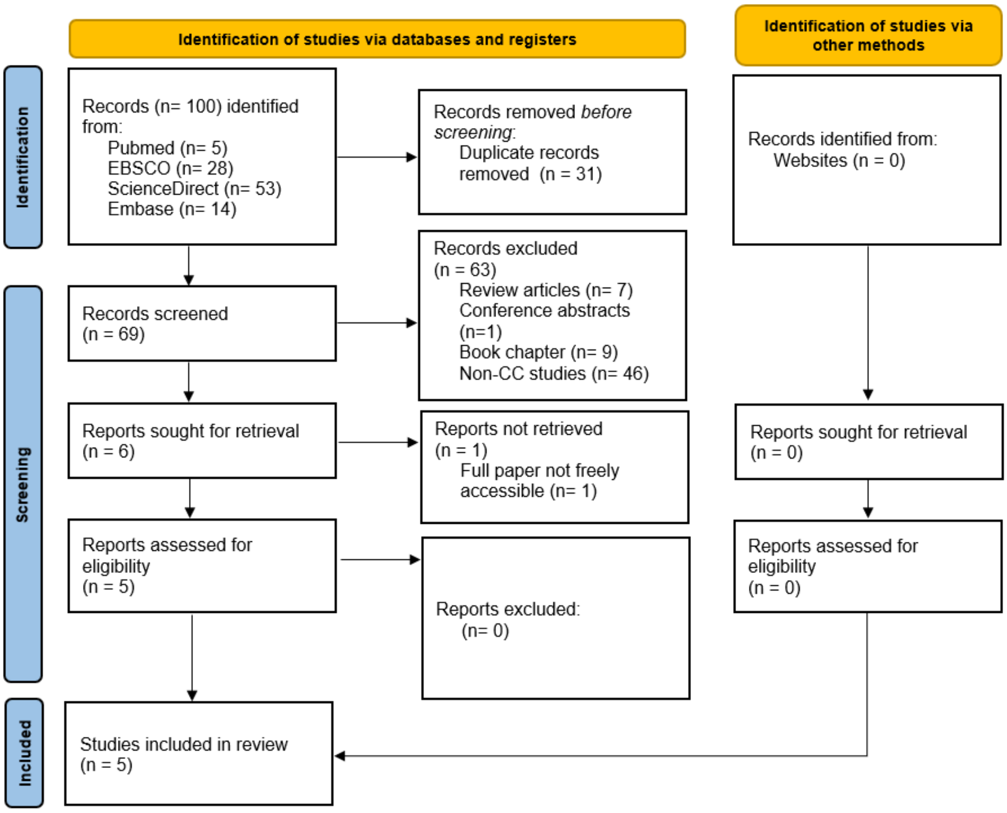
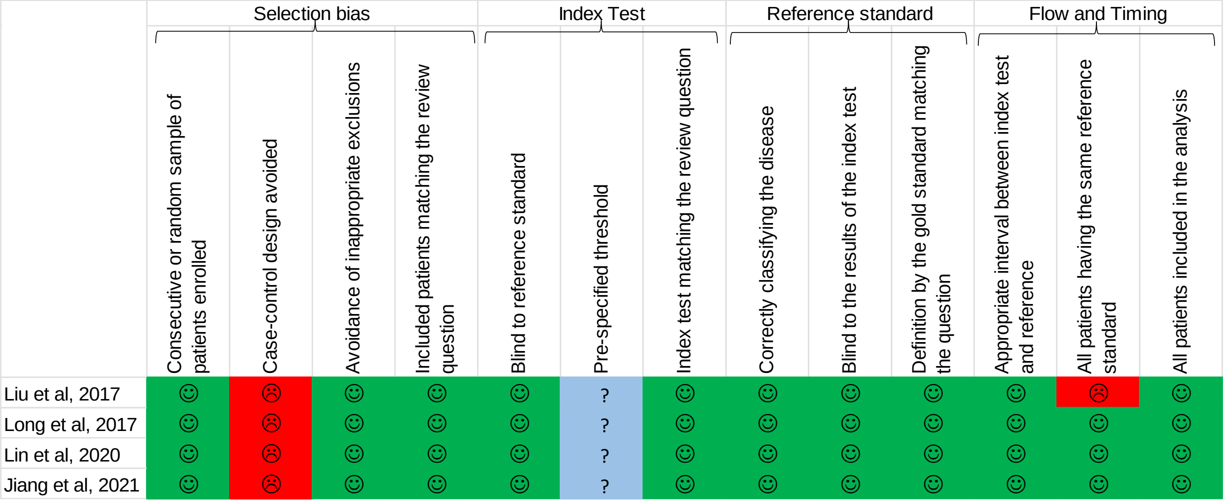

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records (n= 100) identified from:
Pubmed (n= 5)
EBSCO (n= 28)
ScienceDirect (n= 53)
Embase (n= 14)
Records removed before screening:
Duplicate records removed (n = 31)
Records identified from:
Websites (n = 0)
Screening
Records screened (n = 69)
Records excluded (n = 63)
Review articles (n= 7)
Conference abstracts (n=1)
Book chapter (n= 9)
Non-CC studies (n= 46)
Reports sought for retrieval (n = 6)
Reports not retrieved (n = 1)
Full paper not freely accessible (n= 1)
Reports sought for retrieval (n = 0)
Reports assessed for eligibility (n = 5)
Reports excluded: (n= 0)
Reports assessed for eligibility (n = 0)
Included
Studies included in review (n = 5)

**Figure 1. Outline of study selection**

The risk of bias of the retrospective studies[18–21] were assessed using the QUADAS-2[15] tool and the results were described in Figure 2. Overall, the scores ranged from 76.9%[19] to 84.6%[18,20,21]. While, the RCT[22] attained 92% on the 25 items CONSORT-AI[16] checklist, falling short on stating the inclusion and exclusion of the input data and the performance errors analysis (checklist not included).



| | Selection bias | | | | Index Test | | | Reference standard | | | Flow and Timing | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Consecutive or random sample of patients enrolled | Case-control design avoided | Avoidance of inappropriate exclusions | Included patients matching the review question | Blind to reference standard | Pre-specified threshold | Index test matching the review question | Correctly classifying the disease | Blind to the results of the index test | Definition by the gold standard matching the question | Appropriate interval between index test and reference | All patients having the same reference standard | All patients included in the analysis |
| Liu et al, 2017 | ☺ | ☹ | ☺ | ☺ | ☺ | ? | ☺ | ☺ | ☺ | ☺ | ☺ | ☹ | ☺ |
| Long et al, 2017 | ☺ | ☹ | ☺ | ☺ | ☺ | ? | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ |
| Lin et al, 2020 | ☺ | ☹ | ☺ | ☺ | ☺ | ? | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ |
| Jiang et al, 2021 | ☺ | ☹ | ☺ | ☺ | ☺ | ? | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ |

**Figure 2. Bias assessment using QUADAS-2 on the included studies**

General characteristics of the six studies conducted in China between 2017 and 2021 were included in Table 2. Five studies were retrospective[19–22], while one was a randomized controlled trial[18]. The studies focused on various objectives, including CC detection, classification, and prediction. Different approaches were used, such as using slit lamp images for detection and classification,[18,19,21,22] and utilizing risk factors for prediction.[20] The number of images or patients in each study ranged from 350 to 2005, and participant ages ranged from 18.96 to 78.96 months.

**Table 2 General characteristics of the studies included**

| Authors | Year, Data Source(s) | Study type | Country | Brief Description | Study Objectives | Characteristics of population | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | N | Mean age (SD), months | Sex, %male |
| **Liu et al** | 2017; CCPMOH | Retrospective study | China | Training CNN on slit lamp images to diagnose and grade CC | Diagnosis: Detection and classification | 886 images | NR | NR |
| **Long et al** | 2017, CCPMOH, websites, other eye hospitals | Retrospective study | China | Training CNN on slit lamp images to diagnose and grade CC | Diagnosis: Detection and classification | 1349 images | NR | NR |
| **Lin et al** | 2019; multiple eye centers | Randomised controlled trial | China | Comparing CNN and senior consultants on diagnosis and grading of CC | Diagnosis: Detection and classification | 350 images | 78.96 (5.4) | 44 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Lin et al** | 2020; CCPMOH | Case-control retrospective study | China | Training CNN on 11 risk factors based on birth conditions, family medical history, and family environmental factors to predict CC | Prediction | 2005 patients | 37.95 (29.81) | 57.6 |
| **Jiang et al** | 2021; CCPMOH, websites | Retrospective study | China | Training CNN on slit lamp images to grade CC | Diagnosis: Classification | 681 images | 18.96 (10.61) | NR |

CCPMOH: Childhood Cataract Program of the Chinese Ministry of Health; CNN: convoluted neural networks; CC: congenital cataract; SD: standard deviation; NR: not recorded

Table 3 provides an overview of the CNN algorithms used in the reviewed studies. Different CNN models, including AlexNet, Random Forest (RF), Adaptive boosting (Ada), and CCNN ensemble, were employed, paired with classifiers like Support Vector Machine (SVM) and SoftMax.[18–22] The evaluation methods involved cross-validation (4-fold or 5-fold) as well as external validation. Training and external validation datasets consisted of images or patients' data. The study by Lin et al[22], being an RCT, did not have a specific training dataset as they employed CNN which had been validated previously. Thus, the dataset fed onto the CNN was regarded as the external validation dataset. In Liu et al[19], the CNN was not externally validated. Ophthalmologists or cataract experts served as the reference standard for evaluating the algorithm's accuracy. The classification tasks focused on detecting the presence of CC, grading CC based on opacity area, density, and location, as well as predicting CC based on risk factors

**Table 3 Details on algorithm used in each study**

| Studies | CNN Architecture | Classification | Training Dataset | Internal Validation Datasets | External Validation Datasets | Reference Standard |
|---|---|---|---|---|---|---|
| **Liu et al, 2017** | AlexNet + SVM | Presence of cataract; three grades degrees: opacity area, density, location | 886 images | 25% (4-fold cross validation) | No | Panel of 3 ophthalmologists |
| **Long et al, 2017** | AlexNet + Softmax | Presence of cataract; three grades degrees: opacity area, density, location | 1296 images | 20% (5-fold cross validation) | 57 clinical images + 53 website images | Panel of 3 ophthalmologists |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Lin et al, 2019** | AlexNet + Softmax | Presence of cataract; three grades degrees: opacity area, density, location | N/A | N/A | 350 images | Panel of 3 Cataract experts (10 years experience) |
| **Lin et al, 2020** | RF and Ada | Presence of cataract | 1738 patients | 25% (4-fold cross validation) | 267 patients | Panel of 2 ophthalmologists |
| **Jiang et al, 2021** | CCNN-ensemble | Three grades degrees: opacity area, density, location | 470 images | 20% (5-fold cross validation) | 132 clinical images + 79 website images | Panel of 3 ophthalmologists |

SVM: Support Vector Machine; RF: random forest; Ada: adaptive boosting; CCNN-ensemble: ensemble of cost-sensitive CNNs; N/A: not available

Table 4 summarizes diagnostic values in the reviewed studies for identifying CC. Four studies reported varying performance metrics using different CNN architectures.[18–20,22] Accuracy ranged from 87.4% to 98.87%, with variable sensitivity and specificity. AlexNet with Softmax classifier showed relatively high performance[18,22], while RF and Ada algorithms had lower but still decent values.[20] Compared to humans, reported in Long et al[18], CNN algorithm detected all CC cases correctly while all ophthalmologists had one misidentification case due to lighting. In Lin et al[22], CNN sensitivity, specificity, and accuracy were 89.7%, 86.4%, and 87.4% respectively, while senior consultants achieved higher values of 98.4%, 99.6%, and 99.1% respectively based on expert standards.

**Table 4. DL performance on identification of CC**

| | | **Identification** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | | | **Sensitivity** | | | **Specificity** | | | **AUC** | | |
| **Authors** | **CNN Architecture** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** |
| **Liu et al, 2017** | AlexNet + SVM | 97.07 (0.01) | NR | NR | 97.28 (0.01) | NR | NR | 96.83 (0.02) | NR | NR | 0.9686 | NR | NR |
| **Lin et al, 2019** | AlexNet + Softmax | NR | 87.4 | NR | NR | 89.7 | NR | NR | 86.4 | NR | NR | NR | NR |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Long et al 2017** | AlexNet + Softmax | 98.87 | 98.25 | 92.45 | 98.78 | 100 | 100 | 98.95 | 97.67 | 71.43 | 0.9996 | 1 | 0.9232 |
| **Lin et al, 2020** | RF | 81*/79$^+$ | 86*/86$^+$ | NR | 79*/56$^+$ | 80*/58$^+$ | NR | 82*/92$^+$ | 91*/98$^+$ | NR | 0.91*/0.82$^+$ | 0.995*/0.85$^+$ | NR |
| | Ada | 79*/75$^+$ | 85*/85$^+$ | NR | 78*/70$^+$ | 77*/58$^+$ | NR | 81*/78$^+$ | 90*/97$^+$ | NR | 0.89*/0.8$^+$ | 0.91*/0.86$^+$ | NR |
| **Jiang et al, 2021** | CCNN ensemble | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |

*bilateral prediction; $^+$unilateral prediction; NR: not recorded; mean (standard deviation); AUC: area under the curve

Table 5, 6, and 7 present diagnostic parameters of CNNs for grading CC based on area, density, and location.[18,19,21,22] Overall, CNN architectures (AlexNet + Softmax) showed promising results, with Long et al[18] achieving the highest accuracy, sensitivity, and specificity. Generally, performance decreased from training to external datasets across these variables. When compared to human counter parts, there were inconsistent results. In Long et al[18], CNN outperformed ophthalmologists in both classification and grading tasks. In Lin et al[22], AI group achieved lower accuracy in grading CC (90.6% area, 80.2% density, 77.1% location) compared to senior consultants (93.3%, 85.0%, 87.5% respectively).

**Table 5. DL performance on area grading of CC**

| **Authors** | **CNN Architecture** | **Grading - Area** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | | | **Sensitivity** | | | **Specificity** | | | **AUC** | | |
| | | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** |
| **Liu et al, 2017** | AlexNet + SVM | 89.02 (0.01) | NR | NR | 86.63 (0.06) | NR | NR | 90.75(0.04) | NR | NR | 0.98923 | NR | NR |
| **Lin et al,** | AlexNet + Softmax | NR | 90.6 | NR | NR | 91.3 | NR | NR | 88.9 | NR | NR | NR | NR |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2019** | | | | | | | | | | | | |
| **Long et al 2017** | Alex Net + Soft max | 93.98 | 100 | 94.87 | 95.83 | 100 | 90.48 | 91.43 | 100 | 100 | 0.9738 | 1 | 0.9603 |
| **Lin et al, 2020** | RF | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| | Ada | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| **Jiang et al, 2021** | CCNN ensemble | 92.13 | 94.7 | 89.87 | 92.31 | 90.24 | 90 | 92 | 96.7 | 89.47 | 0.9776 | 0.9694 | 0.9465 |

NR: not recorded; mean (standard deviation); AUC: area under the curve

**Table 6. DL performance on density grading of CC**

| **Authors** | **Cnn Architecture** | **Grading - Density** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | | | **Sensitivity** | | | **Specificity** | | | **AUC** | | |
| | | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** |
| **Liu et al, 2017** | Alex Net + SVM | 92.68 (0.01) | NR | NR | 91.05 (0.02) | NR | NR | 93.94 (0.02) | NR | NR | 0.97433 | NR | NR |
| **Lin et al, 2019** | Alex Net + Softmax | NR | 80.2 | NR | NR | 85.3 | NR | NR | 67.9 | NR | NR | NR | NR |
| **Long et al 2017** | Alex Net + Softmax | 95.06 | 92.86 | 84.62 | 95.65 | 85.71 | 90 | 94.29 | 100 | 78.95 | 0.9882 | 1 | 0.9632 |
| **Lin et al, 2020** | RF | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| | Ada | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| **Jiang** | CCNN | 92.77 | 93.18 | 88.61 | 91.43 | 89.29 | 88.52 | 93.85 | 94.23 | 88.89 | 0.9801 | 0.977 | 0.9563 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **et al, 2021** | ensemble | | | | | | | | | | | | |

NR: not recorded; mean (standard deviation); AUC: area under the curve

**Table 7. DL performance on location grading of CC**

| | | **Grading - Location** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | | | **Sensitivity** | | | **Specificity** | | | **AUC** | | |
| **Authors** | **Cnn Architecture** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** | **Training** | **Independent Testing Dataset** | **Website-Based Dataset** |
| **Liu et al, 2017** | AlexNet + SVM | 89.28 (0.03) | NR | NR | 82.70 (0.06) | NR | NR | 93.08 (0.04) | NR | NR | 0.95911 | NR | NR |
| **Lin et al, 2019** | AlexNet + Softmax | NR | 77.1 | NR | NR | 84.2 | NR | NR | 50 | NR | NR | NR | NR |
| **Long et al 2017** | AlexNet + Softmax | 95.12 | 100 | 94.87 | 92.31 | 100 | 100 | 100 | 100 | 91.67 | 0.9808 | 1 | 0.9861 |
| **Lin et al, 2020** | RF | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| | Ada | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR | NR |
| **Jiang et al, 2021** | CCNN ensemble | 92.76 | 93.18 | 87.34 | 89.29 | 90.63 | 87.3 | 95.25 | 94 | 87.5 | 0.9729 | 0.9813 | 0.9306 |

NR: not recorded; mean (standard deviation); AUC: area under the curve

## DISCUSSION

The findings of this study indicate that DL algorithms, especially CNN, for diagnosing and predicting CC demonstrated reasonably high diagnostic values. All accuracy values based on AUC were more than 0.9, which is considered high.[23] CNN's high performance in image analysis can be attributed to its hierarchical feature extraction capabilities, allowing it to extract low-level and gradually learn higher-level features, enhancing recognition performance.[24]

Additionally, CNN's non-linear mapping ability enables it to handle complex and non-linear patterns in images, surpassing the limitations of linear models and traditional machine learning approaches.[25] Moreover, transfer learning empowers CNNs to leverage pre-trained models and fine-tune them on specific datasets, leading to faster convergence and improved recognition performance, particularly when labeled training data is scarce.[26] Although DL algorithms may seem to perform better than the human counterparts in several cases, several methodological deficiencies were identified across the included studies which might result in biases.

There were variabilities in the outcomes of the algorithms, with some focusing on detecting the presence[18,19,22] and grading of CC[18,19,21,22] while other focused on predicting CC[20]. Additionally, it is important to note that all the studies included in our review were conducted in China and utilized datasets consisting only of Chinese patients. This limited geographical and ethnic representation may affect the generalizability of DL algorithms to other demographics and populations. In addition, all studies included in this analysis relied on a reference standard established by an average of 2 to 3 experts. The phenotypes of CC are diverse and abundant, exhibiting a wide range of morphological patterns and varying degrees of severity. This variability can introduce subjectivity when grading CC.[27] Therefore, it is important to exercise caution when interpreting results from studies that rely on a small number of expert graders, as there could be grader bias.

The included studies exhibited variations in the number of data used for training the algorithms, ranging from 350 to 2005. CNNs improve diagnostic accuracy by minimizing the error between their output and the actual image diagnosis. Larger datasets are expected to yield more reliable diagnostic results compared to smaller datasets,[28] as smaller datasets may lead to overfitting where the algorithm memorizes irrelevant noise instead of meaningful patterns.[24] However, it is hard to collect samples given the rarity of the disease[3] Only two studies[20,22] included sample size calculation, indicating a need for more rigorous study design. Despite the challenges in conducting sample size calculations for AI algorithms, it remains an essential aspect of study design and hence should be addressed in future studies.[29]

Several other methodological deficiencies included the omission of poor-quality images and the absence of external validation. Notably, none of the studies provided information regarding the exclusion of poor-quality images, and one study failed to conduct external validation.[19] It is crucial to consider image quality and external validation as significant factors when evaluating algorithm performance in clinical settings.[28]

There are several limitations to our study. The datasets used were homogenous as they were collected in one country. There were also heterogeneities due to differences in imaging modes and internal features of DL models. Accuracy measurements were unavailable for some studies or subdatasets. Some studies lacked external validation or comparison with other professionals.

## CONCLUSION

DL algorithms, particularly those utilizing CNN, exhibit superior performance in diagnosing and predicting CC, outperforming human experts in some cases. Sustained high-quality research is essential to effectively integrate this transformative technology and ultimately reduce visual impairment and blindness associated with CC.

## REFERENCES

1. Sheeladevi S, Lawrenson JG, Fielder AR, Suttle CM. Global prevalence of childhood cataract: a systematic review. Eye (Lond). 2016 Sep;30(9):1160–9.
2. Stevens GA, White RA, Flaxman SR, Price H, Jonas JB, Keeffe J, et al. Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990-2010. Ophthalmology. 2013 Dec;120(12):2377–84.
3. Wu X, Long E, Lin H, Liu Y. Prevalence and epidemiological characteristics of congenital cataract: a systematic review and meta-analysis. Sci Rep. 2016 Jun 23;6:28564.
4. Self JE, Taylor R, Solebo AL, Biswas S, Parulekar M, Dev Borman A, et al. Cataract management in children: a review of the literature and current practice across five large UK centres. Eye (Lond). 2020 Dec;34(12):2197–218.
5. Katre D, Selukar K. The Prevalence of Cataract in Children. Cureus. 2022 Oct;14(10):e30135.
6. You C, Wu X, Zhang Y, Dai Y, Huang Y, Xie L. Visual impairment and delay in presentation for surgery in chinese pediatric patients with cataract. Ophthalmology. 2011 Jan;118(1):17–23.
7. Sheeladevi S, Lawrenson JG, Fielder A, Kekunnaya R, Ali R, Borah RR, et al. Delay in presentation to hospital for childhood cataract surgery in India. Eye (Lond). 2018 Dec;32(12):1811–8.
8. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962 Jan;160(1):106–54.
9. Wu X, Liu L, Zhao L, Guo C, Li R, Wang T, et al. Application of artificial intelligence in anterior segment ophthalmic diseases: diversity and standardization. Ann Transl Med. 2020 Jun;8(11):714.
10. Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. BioMed Eng OnLine. 2019 Dec;18(1):29.
11. Raju M, Pagidimarri V, Barreto R, Kadam A, Kasivajjala V, Aswath A. Development of a Deep Learning Algorithm for Automatic Diagnosis of Diabetic Retinopathy. Stud Health Technol Inform. 2017;245:559–63.
12. ei D, Gong Y, Kang H, Zhang C, Guo Q. Accurate and rapid screening model for potential diabetes mellitus. BMC Med Inform Decis Mak. 2019 Dec;19(1):41.
13. Huang ML, Chen HY. Glaucoma classification model based on GDx VCC measured parameters by decision tree. J Med Syst. 2010 Dec;34(6):1141–7.
14. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association. 2020 Dec 9;27(12):2011–5.
15. Whiting PF, Rutjes AWS, 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 Oct 18;155(8):529–36.
16. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.
17. 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 Apr;88:105906.
18. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng. 2017 Jan 30;1(2):0024.

19. Liu X, Jiang J, Zhang K, Long E, Cui J, Zhu M, et al. Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. PloS one. 2017 Mar 17;12(3):e0168606.
20. Lin D, Chen J, Lin Z, Li X, Zhang K, Wu X, et al. A practical model for the identification of congenital cataracts using machine learning. EBioMedicine. 2020 Jan;51:102621.
21. Jiang J, Wang L, Fu H, Long E, Sun Y, Li R, et al. Automatic classification of heterogeneous slit-illumination images using an ensemble of cost-sensitive convolutional neural networks. Annals of translational medicine. 2021 Apr;9(7):550.
22. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine. 2019 Mar 17;9:52–9.
23. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010 Sep;5(9):1315–6.
24. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug;9(4):611–29.
25. Zhang CL, Wu J. Improving CNN linear layers with power mean non-linearity. Pattern Recognition. 2019 May;89:12–21.
26. Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. Int J Multimed Info Retr. 2022 Mar;11(1):19–38.
27. Reis LM, Semina EV. Genetic landscape of isolated pediatric cataracts: extreme heterogeneity and variable inheritance patterns within genes. Hum Genet. 2019 Sep;138(8–9):847–63.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44.
29. Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics. 2023 Feb 14;24(1):48.