• Review Article •

Deep learning applications for diabetic retinopathy and retinopathy of prematurity diseases diagnosis: a systematic review

Elizabeth Ndunge Mutua[1], *Bernard Shibwabo Kasamani*[1], *Christoph Reich*[2]

[1]School of Computing & Engineering Sciences, Strathmore University, Nairobi 00100, Kenya
[2]Institute for Data Science, Cloud Computing and IT Security, Furtwangen University, Furtwangen 78120, Germany
Correspondence to: Elizabeth Ndunge Mutua. School of Computing & Engineering Sciences, Strathmore University, Nairobi 00100, Kenya. elizabeth.mutua@strathmore.edu
Received: 2024-04-18 Accepted: 2025-04-09

Abstract

● To review the existing deep learning applications for diagnosing diabetic retinopathy and retinopathy of prematurity diseases, the available public retinal databases for the diseases and apply the International Journal of Medical Informatics (IJMEDI) checklist were assessed the quality of included studies; an in-depth literature search in Scopus, Web of Science, IEEE and ACM databases targeting articles from inception up to 31st January 2023 was done by two independent reviewers. In the review, 26 out of 1476 articles with a total of 36 models were included. Data size and model validation were found to be challenges for most studies. Deep learning models are gaining focus in the development of medical diagnosis tools and applications. However, there seems to be a critical issue with most of the studies being published, with some not including information about data sources and data sizes which is important for their performance verification.

● **KEYWORDS:** diabetic retinopathy; retinopathy of prematurity; retinal vessel segmentation; retinal database; deep learning

DOI:10.18240/ijo.2025.08.23

Citation: Mutua EN, Kasamani BS, Reich C. Deep learning applications for diabetic retinopathy and retinopathy of prematurity diseases diagnosis: a systematic review. *Int J Ophthalmol* 2025;18(8):1594-1602

INTRODUCTION

Diabetic retinopathy (DR) is a retina vessel disorder which occurs because of prolonged effects of diabetes mellitus[1]. In the year 2020, people suffering from the disease globally were estimated to be 103.12 million. These numbers are estimated to rise to 160.50 million by 2045, with low- and middle-income countries being affected the most[2]. It is estimated that the people living with DR the statistics will rise to 643 million by 2023 and worse to 783 million by 2045[3]. Retinopathy of prematurity (ROP) is a disease of retinal vascular and capillary proliferation which affects babies born preterm[4]. ROP disease is the leading cause of preventable global childhood blindness[4-6]. ROP disease has five stages of development, ranging from stage 1 to stage 5[7-9]. For stage 1, which is the initial stage, the eye develops a thin demarcation white line separating retinal regions of the eye and hence preventing flow of blood to the rest of the eye[8-9]. This occurs because of abnormal growth of blood vessels and if not diagnosed and treated, the demarcation line grows thicker forming a pinkish ridgeline which now becomes stage 2[8-9]. In stage 3, the pinkish ridgeline grows broader resulting in the formation of abnormal blood vessels. This stage usually requires urgent treatment to prevent progression to the next stages (stage 4 and stage 5)[8-9]. In stage 4, which is an advanced stage, the retina detaches partially[8-9]. For stage 5, the retina detaches completely resulting in blindness. Stages 1 and 2 of ROP can heal without treatment but once the disease progresses to stage 3, it must be diagnosed and treated to stop it from progressing to stages 4 and 5 which are more severe stages[8-9].

Recently, there have been advancements in the development of deep learning (DL) applications for retinal diseases diagnosis. These systems have the capability to utilize huge datasets for training and testing. Even though these models produce promising accuracy, there are observed variations across the studies in the dimensions of model design, results, data sizes and validation methods. The aim of this systematic review was to: 1) identify studies which have used DL for DR and ROP diseases diagnosis, 2) use the IJMEDI guidelines to assess the quality of the included literature, 3) summarize the work of different studies which have applied DL for DR and ROP disease diagnosis, establishing their models design and

performance, 4) summarize the current challenges associated with developing DL models for DR and ROP disease diagnosis for future reference.

MATERIALS AND METHODS

The protocol was registered in International Prospective Register of Systematic Reviews (PROSPERO), an internationally recognized register for systematic reviews under (ID-409975, 21 March 2023) and was approved by Strathmore University Institutional Scientific and Ethical Review Committee (SU-ISERC), certificate number: SU-ISERC1534/22 and additionally cleared by the Kenya National Commission for Science, Technology and Innovation (NACOSTI), license number: NACOSTI/P/23/23702. We confirm that the work did not involve direct participation of human subjects or animals. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[10] were adopted while conducting the systematic review and the study period was from inception to 31st January 2023. The review was done by two independent reviewers. The study questions for this review were structured as "Which deep learning models have been developed for diabetic retinopathy and retinopathy of prematurity diseases diagnosis".

Search Strategy

Electronic database search PRISMA guidelines were applied in conducting the review through evaluating the effects of new interventions, paying much focus to the study objectives other than evaluating the stated use of the interventions. An in-depth literature search was conducted in the following electronic databases: Scopus, Web of Science Core Collection, IEEE, and ACM. The in-depth search was conducted on the databases from inception up to 31st January 2023. Studies reviewed were publications on DL application for DR and ROP. The search terms included a wide range of terms such as: "Diabetes"[MeSH] OR "diabetes"[tiab] OR "diabetes retinopathy"[tiab] OR "Retinopathy"[MeSH] OR "retinopathy"[tiab] OR "retinopathy of prematurity"[tiab] OR "retina image segmentation"[tiab] OR "Deep Learning"[MeSH] OR "Neural Network"[MeSH] OR "diagnosis"[tiab] OR "retinal disease diagnosis" OR "deep learning architecture"[tiab] OR "retina image analysis"[tiab].

Extra manual search In addition to reviewing studies from online databases, we checked for articles related to those on our list of references, and further for more literature from OpenSIGLE.OAI from inception to 31st January 2023. During the search, we identified more articles which were important for the study.

Criteria for studies exclusion and inclusion Studies that were included met the following criteria: 1) DL applications for Diabetic Retinopathy disease diagnosis and 2) DL application for ROP disease diagnosis. Details of importance included year of publication, database(s) used, model design, model accuracy and model validation methods. Publications/studies whose full content access was not granted, reviews and articles whose data sources was not available, studies not related to humans, or any publication not written in English were excluded.

Data Analysis and Synthesis Two reviewers (Kasamani BS and Reich C) first checked the articles titles and abstracts to confirm if they met the study objectives. Studies meeting the inclusion criteria were examined blindly to confirm if they contained methodical elements and results. Information of importance to our study was extracted and reports of the two reviewers were compared. Any disagreements were resolved by consensus. Information recorded included: 1) year of publication, 2) DL architecture used to build the model, 3) results, 4) model limitations, 5) model accuracy, specificity, sensitivity, 6) model validation method, 7) retina database size, description, and owners.

Studies Quality Assessment and Reporting Two reviewers (Kasamani BS and Reich C) used the IJMEDI checklist[11] to assess the quality of included studies. Differences between their reports were solved through discussion. IJMEDI checklist is a quality assessment tool developed for authors and reviewers to assess the quality of their studies in the field of machine learning applications for medical solutions. The tool was developed to assist in distinguishing between high quality studies from mere machine learning applications in the medical field. IJMEDI checklist provides directive for study quality assessment using six dimensions: problem understanding, data understanding, data preparation, modelling, validation, and deployment. The checklist provides a series of thirty questions, each question to be answered as OK (adequately addressed), mR (sufficient but requires improvement) and MR (inadequately addressed). As guided by previous studies[12-13], we associated the values 2, 1, and 0 to be scored for OK, mR, and MR respectively for high-priority items, and 1, 0.5, and 0 to be scored for low-priority items, the maximum number of points being 50. The study quality was divided into low (0-19.5), medium (20-34.5), and high (35-50).

RESULTS

In this section, we summarize the studies containing the development of DL applications for DR or ROP diseases diagnosis and the available retinal databases.

Search Results Figure 1 summarized the procedure for study selection. A total of 1474 studies were identified after the initial search, 2 studies were included after a manual search, making the total included studies to be 1476. After removal of duplicates, 40 studies were identified for further screening. Fourteen studies were excluded for various reasons leading to a total of 26 studies being included in the systematic review.

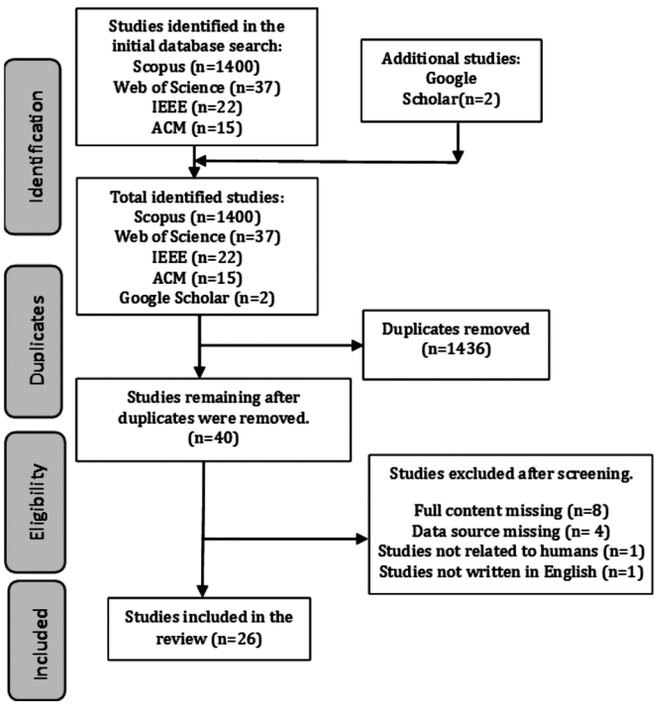

Figure 1 Flow diagram of study selection.

Table 1 summarized studies on DL applications for DR and ROP diseases diagnosis. Fourteen studies with a total of 24 models reported on DL application for DR disease diagnosis, 12 studies with a total of 12 models reported on DL application for ROP disease diagnosis.

Deep Learning Application for Diabetic Retinopathy Disease Diagnosis

Binary classification This section presents a review of studies whose aim was to detect the presence, absence or severity of DR. A study[14] developed an eight layered convolutional neural network (CNN), images were resized achieving an accuracy of 94.5%. A study[15] trained three CNNs using images from three different databases, 88 702 images from Kaggle database, 89 images from DiaretDB1[16] and 107 799 images from E-ophtha database. The work aimed at classifying the presence of the disease (referable: moderate or severe stage) or (non-referable: mild or no disease). A study[17] trained a ResNet named ResNet34[18] using 35 000 images from the Kaggle[19] database to classify images as either with DR, or normal. Image pre-processing techniques applied to improve quality in the study included: Gaussian filter, weighted addition, and normalization. A study[20] developed a 16-layer CNN architecture similar to a VGG-16[21] to determine the presence of referable DR, achieving a sensitivity of 98.2%. The study[22] integrated three CNNs namely Inception V3[23], Inception-ResNet-V2[18] and ResNet 152[24]. The work aimed at classifying images as referable DR or non-referable DR. The model attained an accuracy of 88.21% and area under curve (AUC) of 0.946.

A study[25] constructed a 105 layered CNN named WP-CNN to classify DR as referable and non-referable. The study collected over 60 000 images plus data from STARE dataset[26]. Their model had a better performance than the three models and achieved an accuracy of 94.23% with their data and 90.84% with data from the STARE database. A study[27] applied random forest classification technique to advance the LeNet architecture with an aim of detecting red lesions for DR detection and. achieved a sensitivity of 48.71%. A study[28] customized CNN to detect hard exudate lesions using data from E-ophtha[29] database and HEI-MED[30] database. The model achieved a sensitivity of 0.8990, AUC of 0.9644 using the E-ophtha dataset and a sensitivity of 0.9477, AUC of 0.9323 using HEI-MED dataset.

Multi-level classification of DR This section reviews studies aimed at classifying the different stages of DR. A study[31] developed a CNN model to detect DR and diabetic macular edema (DME). The work used 1748 images from the Messidor-2[32] database and 9963 images from the eyepacs-1 database obtaining a 96.1% sensitivity for the Messidor-2 database, and a 97.5% sensitivity for the eyepacs-1 dataset. This work was limited, particularly, it did not classify the stages of DR or identify the images without DR. A study[33] developed a customized CNN to classify the five stages of DR. A total of 80 000 images from the Kaggle database[19] were used to build the model. Color normalization and resizing was done to achieve 512×512 pixels. The model contained ten convolutional layers, eight maximum pooling layers, and three fully connected layers. The soft maximum function was applied as a classifier, L2 regularization and drop out methods were applied to reduce model overfitting. The model achieved a specificity of 95%, an accuracy of 75% and a sensitivity of 30%. This model had challenges of using images from one database for model building and validation and could not detect lesions on the images. The study[34] developed three models: a deep neural network (DNN), CNN and a back propagation neural network (BNN). The three models were designed using 2000 images from the Kaggle database and resized to 300×300 pixels. Morphological operations were performed on the images such as: conversion to grayscale and edge detection. The CNN model was a pretrained VGG16[21] with sixteen convolutional layers, four maximum pooling layers and three fully connected layers. The DNN was a simple architecture of three fully connected layers and achieved better results than the CNN and the BNN. One database and a few images were used to build the three models and hence the study could not learn more features from the images.

Vessel segmentation approach This section reviews studies which performed vessel segmentation on the images to extract features for DR diagnosis. The study[35] pretrained DEEPLAB-COCO-LARGEFOV[36] to extract retina blood vessels. Data was obtained from two databases: HRF[37] database, and

Table 1 Characteristics of the 26 studies

Author	Year	Data size	Model type	Disease	AUC (95%CI)/accuracy	Specificity/sensitivity
Wang *et al*[28]	2020	E-ophtha, HEI-MED	CNN	DR	0.9644	-/89.9%
Wu *et al*[39]	2020	STARE, DRIVE, CHASE	CNN	DR	STARE=0.983/95.82%, DRIVE=0.9875/96.72%, CHASE=0.9894/96.88%	-
Liu *et al*[25]	2019	Private database, STARE=60000	ResNet, SeNet, DenseNet	DR	ANN, -/90.84%	ANN
Pires *et al*[20]	2019	Kaggle, Messidor-2	VGG-16	DR	-	-/98.2%
Yan *et al*[27]	2019	DIARETDB1	CNN	DR	-	-/48.71%
Zago *et al*[26]	2020	DIARETDB1, Kaggle, Messidor-2, DDR, IDRiD, DIARETDB0	Customize DCNN, VGG-16	DR	0.912/-	-/94%
Jiang *et al*[22]	2019	DRIVE	Inception V3, Inception-ResNet-V2, ResNet 152	DR	0.946/88.21%	ANN
Esfahani *et al*[17]	2018	Kaggle=35000	ResNet	DR	-/85%	-/86%
Dutta *et al*[34]	2018	Kaggle=2000	DNN, Customize DCNN, BNN, VGG16	DR	-	-
Quellec *et al*[15]	2017	Kaggle=88702, DiaretDB1=89, E-ophtha=107799	AlexNet	DR	Kaggle=0.954/-, E-ophtha=0.949/-	ANN
Xu *et al*[14]	2017	Kaggle=1000	Customize DCNN	DR	ANN, -/94.5%	ANN
Gulshan *et al*[31]	2016	Messidor-2=1748, eyepacs-1=9963	Inception-V3	DR, DME	-	93%, Messidor-2=96.1%, eyepacs
Pratt *et al*[33]	2016	Kaggle=80000	Customize DCNN	DR	-/75%	95%, 30%
Vengalil *et al*[35]	2016	DRIVE, HRF	DEEPLAB-COCO-LARGEFOV	DR	0.894/93.94%	-
Chen *et al*[54]	2021	America and Nepal (private database) total range=2668 to 52249	Customize DCNN	ROP	0.984/-	-
Huang *et al*[47]	2021	China (private database)=10000	Customize DCNN	ROP	0.98/-	91.13%, 95.92%
Wang *et al*[50]	2021	China (private database)=10000	Customize DCNN	ROP	0.98/-	91.13%, 95.92%
Mao *et al*[48]	2020	China (private database) total range=2668 to 52249	Customize DCNN	ROP	0.98/-	91.13%, 95.92%
Tong *et al*[49]	2020	China (private database)=10000	Customize DCNN	ROP	-	91.13%, 95.92%
Yildiz *et al*[53]	2020	America and Mexico (private database) total range=2668 to 52249	Customize DCNN	ROP	0.98/-	-
Hu *et al*[46]	2019	China (private database) total range = 2668 to 52,249	Customize DCNN	ROP	0.984/-	95.72%, 98.15%
Tan *et al*[55]	2019	New Zealand (private database) total range=2668 to 52249	Customize DCNN	ROP	0.98/-	91.13%, 95.92%
Zhang *et al*[51]	2018	China (private database)=10000	Customize DCNN	ROP	0.984/-	95.72%, 98.15%
Brown *et al*[52]	2018	North America (private database) total range=2668 to 52249	Customize DCNN	ROP	0.98/-	91.13%, 95.92%
Wang *et al*[45]	2018	China (private database)=10000	Customize DCNN	ROP	0.984/-	95.72%, 98.15%
Raja Sankari *et al*[44]	2023	Private database	Customize DCNN	ROP	-/94.5%	93%, 94%

AUC: Area under curve; CI: Confidence interval; RBFCN: Region based fully convolutional network; DCNN: Dynamic convolutional neural network; BNN: Bayesian neural network; CNN: Convolutional neural network; ANN: Artificial neural network; DR: Diabetic retinopathy; ROP: Retinopathy of prematurity.

DRIVE[38] database and obtained an accuracy of 93.94% with AUC of 0.894. A study by[39] developed a customized CNN to extract retina blood vessels and used them as patches to detect DR. Data was obtained from STARE[40], DRIVE[38] and CHASE[41] databases achieving an accuracy of 95.82%, 96.72% and 96.88%, and AUC of 98.30%, 98.75% and 98.94% for the DRIVE, STARE and CHASE, respectively. A study[42] did a comparative study to compare the performance of three models: Xception, InceptionV3 and ResNet-50 for DR disease detection and using vessel segmentation. Inception model outperformed the other two models and provided a better classification. A study[43] developed a customized CNN, DenseNet-169 by adding a block attention module for vessel segmentation and disease severity prediction. Another study[44] pretrained Xception model with hierarchical image clustering algorithm to detect patches on the images and grade fundus images for DR disease diagnosis.

Deep Learning Application for Retinopathy of Prematurity Disease Diagnosis

Data preparation This section reviews studies whose data sources and sizes were provided in the study. Sixteen studies were reviewed[45-60], seven of which used data from China[45-51], one got data from North America[52], one from America and Mexico[53], one from America and Nepal[54] and one got data from New Zealand[55]. All studies used data collected between the years 2011 to 2020. Five studies excluded images of poor quality[49-52,55]. Data was augmented before training the models by ten studies[44,46-50,53-55,61]. Eleven studies[46-48,50-51,54-55,57-60]

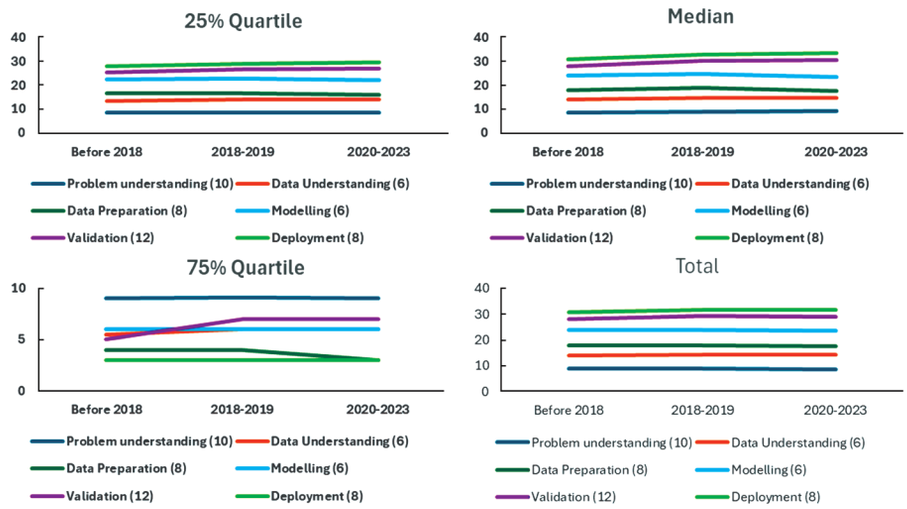

Figure 2 Median and quartile of quality assessment scores according to time of publication.

collected data with a gestational age (GA) of 30.9, birth weight (BW) of 1501.25 g. A total range of 2668 to 52 249 images were collected by all the eleven studies. Seven studies developed algorithms for ROP disease diagnosis[44-47,51,54,60]. Seven algorithms developed algorithms to detect ROP plus disease with a minimum of 5358 images[48-50,53,55-56].

Model design and performance This section reviews studies whose model design and results were included. All the sixteen studies[44-59] customized CNNs architectures ResNet, ImageNet, U-Net, and VGG-16 and applied transfer learning except Huang *et al*[47]. Ten studies used more than 10 000 images[44-45,47,49-51,57-60]. One to five expert graders were used by all studies as a reference standard, with the majority of the graders agreeing upon the images used for the studies. Seven studies conducted external validation of their models[44-45,52-55,60]. Five studies were able to detect the presence of ROP with an average AUC of 0.984[45-47,51,54]. Six studies were able to detect the presence of plus disease with an average sensitivity and specificity of 91.13% and 95.92%, respectively[47-50,52,55]. Three studies reported their model sensitivity and specificity at an average of 95.72% and 98.15% respectively[45-46,51]. One study[51] did a comparison of the model performance with the results of an eye specialist and achieved a sensitivity of 94.1% and specificity of 99.3%. Six studies[47-48,50,52-53,55] produced an average AUC of 0.98. Two studies[48,52] were able to detect the presence of the pre-plus disease at an average sensitivity and specificity of 96.2% and 95.7% respectively.

SUMMARY OF KEY FINDINGS

Included Studies Quality Assessment To the best of our knowledge, this is the first systematic review of studies on DL models for ROP disease diagnosis using the IJMEDI checklist for quality assessment. Figure 2 and Table 2 is used to explain the results. For all studies, it is observable that the median and quartile increased over the years, this is an indication that more studies of good quality are being published over time.

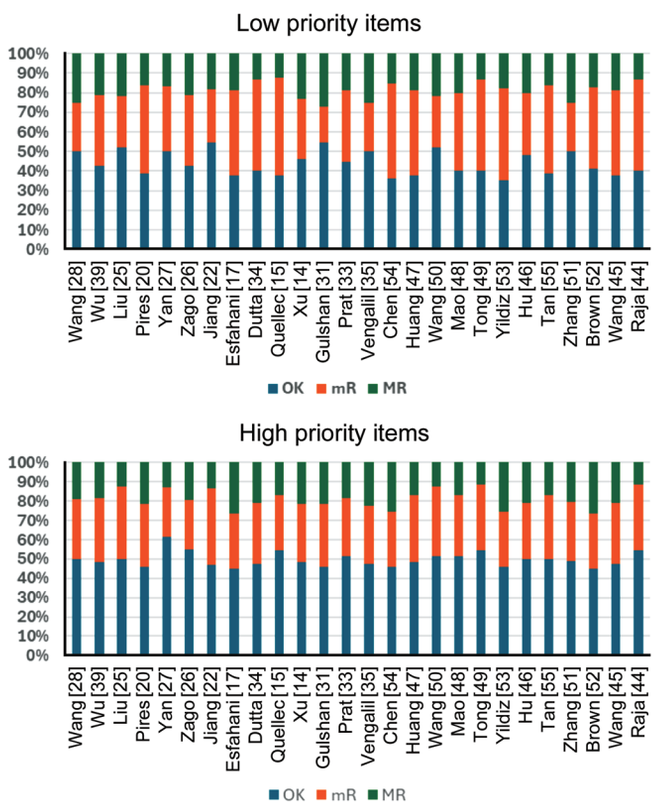

Figure 3 Proportion of different answers in the depth- and low-priority items OK: Adequately addressed; mR: Sufficient but improvable; MR: Inadequately addressed.

In terms of the scores for all answers. For high quality studies with a total score greater than or equal to 32.5, the proportion for OK in high-priority items reached 70% versus their score for OK in low-priority items which was at an average of 30%. Consequently, for low quality studies whose scores were less than or equal to 27, the proportion for the answers MR in high-priority items were more than 30%. Most studies failed to provide important information about the data used as shown in Figure 3. Missing information of importance included: source of data, database size, dataset size used for training the model,

Table 2 Quality assessment scores of the 26 studies according to the IJMEDI checklist

Author	Problem understanding (10)	Data Understanding (6)	Data Preparation (8)	Modelling (6)	Validation (12)	Deployment (8)	Total (50)
Wang *et al*[28]	8	5	3	6	3	3	28
Wu *et al*[39]	8	5.5	3	6	5	3	30.5
Liu *et al*[25]	8	6	2	6	3	2.5	27.5
Pires *et al*[20]	8.5	6	4	6	7	2.5	34
Yan *et al*[27]	9.5	4	2	6	4	2	27.5
Zago *et al*[26]	8.5	4	3	6	5	3	29.5
Jiang *et al*[22]	7	6	2	6	3	2	26
Esfahani *et al*[17]	8.5	5.5	5	6	7	3	35
Dutta *et al*[34]	9	6	4	6	7	2	34
Quellec *et al*[15]	9.5	5	3	6	8	2	33.5
Xu *et al*[14]	9	5.5	4	6	4	3	31.5
Gulshan *et al*[31]	8.5	6	4	6	2	3	29.5
Pratt *et al*[33]	8.5	5	3	6	5	2.5	30
Vengalil *et al*[35]	8.5	5.5	4	6	3	3	30
Chen *et al*[54]	9	5.5	5	6	8	2.5	36
Huang *et al*[47]	8.5	6	3	6	7	3	33.5
Wang *et al*[50]	8.5	6	2	6	3	2.5	28
Mao *et al*[48]	9	5.5	3	6	6	3	32.5
Tong *et al*[49]	9.5	6	2	6	7	2	32.5
Yildiz *et al*[53]	9	5.5	5	6	8	3	36.5
Hu *et al*[46]	9.5	5.5	4	6	4	2.5	31.5
Tan *et al*[55]	9	6	3	6	7	2.5	33.5
Zhang *et al*[51]	9.5	6	4	6	3	3	31.5
Brown *et al*[52]	8.5	5.5	5	6	6	2.5	33.5
Wang *et al*[45]	9	6	4	6	7	3	35
Raja Sankari *et al*[44]	9.5	6	2	6	7	2	32.5

testing and validation which explains the curves for the 25% and 75% quartiles as shown by Figure 2. The scoring rate for the 25%, median and 75% quartile curves was about 40%, 50%, and 60% respectively showing some significant improvement in model design but challenges of model validation. It was also noted that there were more studies published for the years between 2018 and 2019 than all other years. From this review, it is also observable that most studies had high scores with the least being 8 out of 10 for problem understanding and low scores with the highest being 3 out of 8 for deployment. This means that most studies were closed after publishing results while few studies had their output(s) implemented into real-life clinical use. It is also observed that many studies did not include information on the data processing techniques applied. The majority of the studies used data from public databases, while some used few datasets which are deemed sufficient for training and testing their models. It is also noted that Kaggle and HVDROPDB databases are the only publicly available databases containing images of ROP.

CHALLENGES AND FUTURE TRENDS

DL applications have widely been developed to diagnose DR and ROP. Despite the advancements, there are still challenges to be addressed. In this section, we discuss some challenges and future trends for these technologies which could help developers in this field.

Image Quality From this review, we were able to identify the following two common problems associated with developing DL applications for DR and ROP diseases diagnosis. 1) Low quality images: All databases contained colored and non-colored images. Image resolution and clarity vary from colored to non-colored images and this may result in an increase to the prediction errors. There is a need to ensure image quality assessment before utilizing the images for model development. 2) Devices inconsistency: There are many different cameras used to capture retina images. Device inconsistency produces images with different orientation, size, and resolution. Images should be pre-processed to increase resolution, harmonize sizes and orientation.

Model Training DL models require huge datasets for training the models to achieve better results. All databases did not contain adequate images for building DL models. To address this challenge, three methods are proposed: First, transfer

learning can be used to allow models to utilize the knowledge acquired from a previous task to be used for a similar related task which boosts performance of the model classification for the current task. This method also helps to enhance the original input representations of data and for mapping the data. The second approach which can be used is customizing an existing similar model and changing some layers and fine tuning some parameters[45-49]. This is useful, since the model being customized is already trained with enough dataset(s) and images are already resized, therefore, not much additional work is required for data preparation. However, this method is limited in the fact that not every model should be customized especially for bioinformatics solutions. The third approach is applying some simulations to increase data volumes. Simulators can be developed to simulate the number of required images. However, it should be noted that not every problem will use simulators for data generation.

Imbalanced Data Data in many databases are always imbalanced with some sets being more or lesser than the others. Training a DL model with imbalanced data will result in inaccurate results. The following two techniques can be used to solve this problem. First, developers require to utilize the correct criteria for evaluating the model loss and prediction result functions[50]. Another solution to this problem is to utilize the weighted cross entropy loss function[26-27]. This allows the model to produce good results with small classes through enabling the model to up-sample small classes or down-sample large classes.

CONCLUSION

In conclusion, DL models are gaining focus in the development of medical diagnosis tools and applications. However, there seems to be a critical issue with most of the studies being published, with some not including information about data sources and data sizes. Most of the studies reviewed in this paper used data from one database for development, testing and model validation and cannot confirm the validity of their model results. It is also observed that Kaggle and HVDROPDB databases is the only publicly available databases containing images on ROP disease and there is need for the privately owned ROP disease databases to be made public. We also recommend that researchers should work to extend their studies beyond publishing to clinical practice and use. Addressing these issues will promote the development of effective, efficient, and practical DL models and applications for DR and ROP disease diagnosis.

ACKNOWLEDGEMENTS

Foundation: Supported by DAAD, Google Research, and the Organization for Women in Science for the Developing World (OWSD).

Conflicts of Interest: Mutua EN, None; **Kasamani BS,** None; **Reich C,** None.

REFERENCES

1 Shukla UV, Tripathy K. Diabetic retinopathy. 2023 Aug 25. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2025.

2 Teo ZL, Tham YC, Yu M, *et al*. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology* 2021;128(11):1580-1591.

3 Kropp M, Golubnitschaja O, Mazurakova A, *et al*. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. *EPMA J* 2023;14(1):21-42.

4 Wood EH, Chang EY, Beck K, *et al*. 80 Years of vision: preventing blindness from retinopathy of prematurity. *J Perinatol* 2021;41(6): 1216-1224.

5 Braimah IZ, Enweronu-Laryea C, Sackey AH, *et al*. Incidence and risk factors of retinopathy of prematurity in Korle-Bu Teaching Hospital: a baseline prospective study. *BMJ Open* 2020;10(8):e035341.

6 Coyner AS, Chen JS, Singh P, *et al*. Single-examination risk prediction of severe retinopathy of prematurity. *Pediatrics* 2021;148(6):e2021051772.

7 Hong EH, Shin YU, Cho H. Retinopathy of prematurity: a review of epidemiology and current treatment strategies. *Clin Exp Pediatr* 2022;65(3):115-126.

8 Kaur K, Mikes BA. Retinopathy of prematurity. 2025 Jun 2. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2025.

9 Diggikar S, Gurumoorthy P, Trif P, *et al*. Retinopathy of prematurity and neurodevelopmental outcomes in preterm infants: a systematic review and meta-analysis. *Front Pediatr* 2023;11:1055813.

10 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med* 2009;6(7):e1000097.

11 Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform* 2021;153:104510.

12 Kamel Rahimi A, Canfell OJ, Chan W, *et al*. Machine learning models for diabetes management in acute care using electronic medical records: a systematic review. *Int J Med Inform* 2022;162:104758.

13 Zhou Y, Ge YT, Shi XL, *et al*. Machine learning predictive models for acute pancreatitis: a systematic review. *Int J Med Inform* 2022;157:104641.

14 Xu K, Feng D, Mi H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. *Molecules* 2017;22(12):2054.

15 Quellec G, Charrière K, Boudi Y, *et al*. Deep image mining for diabetic retinopathy screening. *Med Image Anal* 2017;39:178-193.

16 Kauppi T, Kalesnykiene V, Kamarainen JK, *et al*. The DIARETDB1 diabetic retinopathy database and evaluation protocol. *Proceedings of the British Machine Vision Conference 2007.* Warwick. British Machine Vision Association, 2007:15.1-15.10.

17 Esfahani M, Ghaderi M, Kafiyeh R. Classification of diabetic and

normal fundus images using new deep learning method. *Leonardo Electron J Pract Technol* 2018;17(32):233-248.

18 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. AAAI Press, 2017;4278-4284.

19 Kaggle dataset. https://kaggle.com/c/diabetic-retinopat hy-detection

20 Pires R, Avila S, Wainer J, *et al*. A data-driven approach to referable diabetic retinopathy detection. *Artif Intell Med* 2019;96:93-106.

21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014:1409.1556. https://arxiv.org/abs/1409.1556v6.

22 Jiang H, Yang K, Gao M, *et al*. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. *Annu Int Conf IEEE Eng Med Biol Soc* 2019;2019:2045-2048.

23 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. IEEE, 2016:2818-2826.

24 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA. IEEE, 2016:770-778.

25 Liu Y, Li Z, Xu C, *et al*. Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network. *Artif Intell Med* 2019;99:101694.

26 Zago GT, Andreão RV, Dorizzi B, *et al*. Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Comput Biol Med* 2020;116:103537.

27 Yan Y, Gong J, Liu Y. A novel deep learning method for red lesions detection using hybrid feature. *2019 Chinese Control and Decision Conference (CCDC)*. Nanchang, China. IEEE, 2019:2287-2292.

28 Wang H, Yuan GH, Zhao XG, *et al*. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. *Comput Methods Programs Biomed* 2020;191:105398.

29 E-Ophtha dataset. http://www.adcis.net/en/Download-ThirdParty/E-Ophtha.h/2013

30 Giancardo L, Meriaudeau F, Karnowski TP, *et al*. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Med Image Anal* 2012;16(1):216-226.

31 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316(22):2402-2410.

32 Decencière E, Zhang XW, Cazuguel G, *et al*. Feedback on a publicly distributed image database: the messidor database. *Image Anal Stereol* 2014;33(3):231.

33 Pratt H, Coenen F, Broadbent DM, *et al*. Convolutional neural networks for diabetic retinopathy. *Procedia Comput Sci* 2016;90:200-205.

34 Dutta S, Manideep BC, Basha SM, *et al*. Classification of diabetic retinopathy images by using deep learning models. *Int J Grid Distrib Comput* 2018;11(1):99-106.

35 Vengalil SK, Sinha N, Kruthiventi SSS, *et al*. Customizing CNNs for blood vessel segmentation from fundus images. *2016 International Conference on Signal Processing and Communications (SPCOM)*. Bangalore, India. IEEE, 2016:1-4.

36 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014:1412.7062. https://arxiv.org/abs/1412.7062v4.

37 Budai A, Bock R, Maier A, *et al*. Robust vessel segmentation in fundus images. *Int J Biomed Imaging* 2013;2013:154860.

38 Staal J, Abràmoff MD, Niemeijer M, *et al*. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans Med Imaging* 2004;23(4):501-509.

39 Wu Y, Xia Y, Song Y, *et al*. NFN+: a novel network followed network for retinal vessel segmentation. *Neural Netw* 2020;126:153-162.

40 Hoover AD, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans Med Imag* 2000;19(3):203-210.

41 Owen CG, Rudnicka AR, Mullen R, *et al*. Measuring retinal vessel tortuosity in 10-year-old children: validation of the Computer-Assisted Image Analysis of the Retina (CAIAR) program. *Invest Ophthalmol Vis Sci* 2009;50(5):2004-2010.

42 Farag MM, Fouad M, Abdel-Hamid AT. Automatic severity classification of diabetic retinopathy based on DenseNet and convolutional block attention module. *IEEE Access* 2022;10:38299-38308.

43 Deepa V, Sathish Kumar C, Cherian T. Automated grading of diabetic retinopathy using CNN with hierarchical clustering of image patches by Siamese network. *Phys Eng Sci Med* 2022;45(2):623-635.

44 Raja Sankari V, Snekhalatha U, Chandrasekaran A, *et al*. Automated diagnosis of Retinopathy of prematurity from retinal images of preterm infants using hybrid deep learning techniques. *Biomed Signal Process Control* 2023;85:104883.

45 Wang J, Ju R, Chen Y, *et al*. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* 2018;35:361-368.

46 Hu J, Chen Y, Zhong J, *et al*. Automated analysis for retinopathy of prematurity by deep neural networks. *IEEE Trans Med Imaging* 2019;38(1):269-279.

47 Huang Y, Basanta H, Kang E, *et al*. Automated detection of early-stage ROP using a deep convolutional neural network. *Br J Ophthalmol* 2021;105(8):1099-1103.

48 Mao J, Luo Y, Liu L, *et al*. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. *Acta Ophthalmol* 2020;98(3):e339-e345.

49 Tong Y, Lu W, Deng Q, *et al*. Automated identification of retinopathy of prematurity by image-based deep learning. *Eye Vis* (*Lond*) 2020;7:40.

50 Wang J, Ji J, Zhang MZ, *et al*. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. *JAMA Netw Open* 2021;4(5):e218758.

51 Zhang Y, Wang L, Wu ZQ, *et al*. Development of an automated

screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. *IEEE Access* 2018;7:10232-10241.

52 Brown JM, Campbell JP, Beers A, *et al*. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol* 2018;136(7):803-810.

53 Yildiz VM, Tian P, Yildiz I, *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(2):10.

54 Chen JS, Coyner AS, Ostmo S, *et al*. Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras. *Ophthalmol Retina* 2021;5(10): 1027-1035.

55 Tan Z, Simkin S, Lai C, *et al*. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. *Transl Vis Sci Technol* 2019;8(6):23.

56 Ramachandran S, Niyas P, Vinekar A, *et al*. A deep learning framework for the detection of Plus disease in retinal fundus images of preterm infants. *Biocybern Biomed Eng* 2021;41(2):362-375.

57 Hanif A, Yıldız İ, Tian P, *et al*. Improved training efficiency for retinopathy of prematurity deep learning models using comparison versus class labels. *Ophthalmol Sci* 2022;2(2):100122.

58 Coyner AS, Chen JS, Chang K, *et al*. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity diagnosis. *Ophthalmol Sci* 2022;2(2):100126.

59 Wu Q, Hu Y, Mo Z, *et al*. Development and validation of a deep learning model to predict the occurrence and severity of retinopathy of prematurity. *JAMA Netw Open* 2022;5(6):e2217447.

60 Feng W, Ju L, Ma T, *et al*. Automatic retinopathy of prematurity staging with a semi-supervised deep learning method based on dual consistency regularization. 2022, Preprint (Version 1). Research Square. https://doi.org/10.21203/rs.3.rs-2154727/v1

61 Bai A, Carty C, Dai S. Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: a systematic review. *Saudi J Ophthalmol* 2022;36(3):296-307.