Received: 31 December 2021 | Revised: 8 May 2022 | Accepted: 3 June 2022

DOI: 10.1049/cvi2.12116

REVIEW

Deep learning in the grading of diabetic retinopathy: A review

Nurul Mirza Afiqah Tajudin[1] | Kuryati Kipli[1] | Muhammad Hamdi Mahmood[2] | Lik Thai Lim[3] | Dayang Azra Awang Mat[1] | Rohana Sapawi[1] | Siti Kudnie Sahari[1] | Kasumawati Lias[1] | Suriati Khartini Jali[4] | Mohammed Enamul Hoque[1]

[1]Department of Electrical and Electronics Engineering, University Malaysia Sarawak (UNIMAS), Sarawak, Malaysia

[2]Department of Para-Clinical Sciences, Faculty of Medicine and Health Sciences (FMHS), University Malaysia Sarawak (UNIMAS), Sarawak, Malaysia

[3]Department of Ophthalmology, Faculty of Medicine and Health Sciences (FMHS), University Malaysia Sarawak (UNIMAS), Sarawak, Malaysia

[4]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia

Correspondence

Kuryati Kipli, Department of Electrical and Electronics Engineering, University Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia.
Email: kkuryati@unimas.my

Funding information

Ministry of Science Technology and Innovation, Malaysia, Grant/Award Number: GL/F02/TeD1/2021 (TDF05211383)

Abstract

Diabetic Retinopathy (DR) grading into different stages of severity continues to remain a challenging issue due to the complexities of the disease. Diabetic Retinopathy grading classifies retinal images to five levels of severity ranging from 0 to 5, which represents No DR, Mild non-proliferative diabetic retinopathy (NPDR), Moderate NPDR, Severe NPDR, and proliferative diabetic retinopathy. With the advancement of Deep Learning, studies on the application of the Convolutional Neural Network (CNN) in DR grading have been on the rise. High accuracy and sensitivity are the desired outcome of these studies. This paper reviewed recently published studies that employed CNN for DR grading to 5 levels of severity. Various approaches are applied in classifying retinal images which are, (i) by training CNN models to learn the features for each grade and (ii) by detecting and segmenting lesions using information about their location such as microaneurysms, exudates, and haemorrhages. Public and private datasets have been utilised by researchers in classifying retinal images for DR. The performance of the CNN models was measured by accuracy, specificity, sensitivity, and area under the curve. The CNN models and their performance varies for every study. More research into the CNN model is necessary for future work to improve model performance in DR grading. The Inception model can be used as a starting point for subsequent research. It will also be necessary to investigate the attributes that the model uses for grading.

1 | INTRODUCTION

Diabetic Retinopathy (DR) screenings require ophthalmologists to evaluate the retinal fundus images and it has become more difficult to offer expert eye care to everyone as the diabetes population grows. However, screening of DR has to be carried out routinely for diabetic patients, which places a huge responsibility on the experts as the growing number of diabetic patients affects their efficiency and causes delays in DR diagnosis and treatments. The increasing gap has initiated the demand for automated DR screening systems and arrangements. With the advancement of technology, automated grading is a solution for DR screening that offers several advantages including increasing efficiency, reproducibility, and scalability, as efficient evaluations of retinal images are needed to support the already substantial manual laborious time-consuming screening work, which can be error-prone. Therefore, there is a need for an automated DR grading system to

Abbreviations: AUC, area under the curve; CNN, Convolutional Neural Network; DL, Deep Learning; DR, Diabetic Retinopathy; NDPR, non-proliferative diabetic retinopathy; PDR, proliferative diabetic retinopathy.

analyse the pattern and characteristics of different DR severity levels that have no subconscious biases nor subjectivity [1].

Researchers have applied different methods to solve diabetes-related problems ranging from detection, classification, and prediction. Throughout the years, more techniques are being developed to automate DR diagnosis [2] such as decision tree, random forest [3, 4] and feature selection [5]. Decision tree and random forest are examples of machine learning techniques that have been widely used in medical diagnosis of Diabetes. Feature selection algorithm is applied for better selection of classifiers to enhance diagnostic accuracy. Published studies on classification for DR detection and grading of its severity level have obtained significant results that may be beneficial in clinical settings in the future. The benefits of automated grading of the DR severity level include enhancing screening programme efficiency and coverage, reducing barriers to access, and improving patient outcomes through early timely detection and treatment [6]. The performance of Deep Learning (DL) algorithms for DR-related problems has been good; however, it is not adequate for optimum and practical clinical deployment due to image-related factors. For example, different image compression and fundus field of view will make a significant impact on the DL model performance [7]. Each step taken before the training process of the model is important; thus, the review will also summarise varieties of methods taken before the classification process. The classification model used in every research differs as their model performance results suggested. The differences in each model will be discussed. The review will be focussing mainly on the various methods used in DR detection and grading, including published work that employed the Convolutional Neural Network (CNN) for DR grading based on the 5 severity levels [8].

The paper is organised into several sections. Section 2 gives an overview of DR and the background of DL for DR analysis. Section 3 describes the different methods based on DL that have been applied in DR detection and grading. Section 4 details the published studies on the application of CNN in DR grading to five severity levels which includes the datasets used, the CNN models applied, and the performance measures for each model. Section 5 presents the challenges faced in applying the models for DR grading. Finally, Section 6 concludes the review studies.

The main contribution of this study is the comparative analysis of different CNN models for DR screening specifically for DR grading. This paper summarised published studies that focussed on DR grading from 2017 to 2022. The studies employed different databases; several CNN models with their performance measured are discussed. Based on the review, the challenges encountered by researchers that utilised CNN models for DR grading were discovered. The challenges highlighted will expose the potential future work in this research area.

2 | BACKGROUND STUDY

Diabetic Retinopathy is a common ocular complication of diabetes that involves retinal abnormalities, which remain a leading cause of visual loss in working-age populations. Physiologically, the retina is a light-sensitive layer, which consists of four main sub-layers, firstly the outer neural layer, containing nerve cells and blood vessels, secondly the photoreceptor layer, a single layer that contains the light-sensing rods and cones, thirdly the pigmented retinal epithelium (PRE) and finally the choroid, consisting of connective tissue and capillaries. The retina is finely insulated from the bloodstream by a barrier known as the blood-retinal barrier (BRB). The outer part of BRB is located at the PRE, which serves to regulate movements of nutrients and solutes through the retinal sublayers. The inner BRB is formed by vascular endothelial of the inner retina and tight junction. Retinal vasculature has a high metabolic demand thus making it susceptible to damage due to oxidative stress, which occurs in pathologic conditions such as chronic diabetes. Diabetic Retinopathy has been recognized as a microvascular disease where retinal abnormalities such as vascular changes, haemorrhages, and fluid extravasation eventually lead to one's vision distortion and reduction [9]. Thus, regular screening and early detection are necessary to prevent DR from worsening before getting treatment.

Clinically, DR is divided into two stages; (i) the early stage of DR is represented by non-proliferative diabetic retinopathy (NPDR), and (ii) the advanced stage of DR is represented by proliferative diabetic retinopathy (PDR) [10]. During the NPDR stage, the main observations in the retinal vasculature focussed on the increased vascular permeability and capillary occlusion wherein microaneurysms, haemorrhages, and hard exudates can be detected by fundus photography although the patients may be asymptotic [10]. Non-proliferative diabetic retinopathy is divided into mild, moderate, and severe. While PDR is defined as neovascularisation or new vessels which are abnormal, which can be classified as new vessels on the optic disc or new vessels elsewhere *n*, especially in tissues where circulation has been impaired by trauma or disease. Based on Table 1, the progression of DR is classified into 5 stages (No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and PDR). The different features in each stage allow the CNN supervised network of DL to be applied for DR classification. Figure 1 shows example of images that represents the 5 severity levels.

Medical diagnosis has gone beyond the manual diagnostics process as time transcends. As computer-aided diagnosis (CAD) become one of the major research subjects in medical imaging, more CAD schemes for various diseases have been developed. Researchers have been combining the idea of CAD, artificial intelligence (AI), and DL for DR in hope of assisting ophthalmologists in grading retinal images.

The rapid growth of AI allows DL, which is a field within AI, to be widely applied in medical imaging analysis. Deep Learning is commonly used for tasks such as detection and classification gave the input data as either labelled or unlabelled data. It compiles trained multi-layer networks of artificial neurons which can automatically identify valuable features such as lines, edges, and shapes [11]. For example, in retinal images, when a retinal vessel is branching out to form abnormal new vessels, it signifies an abnormality in the image. Thus, an analysis of the image is needed to conclude the patient's disease changes affecting the retina. A DL algorithm will analyse the

TABLE 1 International clinical DR disease severity scale (ICDRDSS) adapted from Ref. [8]

Diabetic retinopathy	Findings observable on dilated ophthalmoscopy
No apparent DR	No abnormalities
Mild NPDR	Microaneurysms only
Moderate NPDR	More than just aneurysms, but less than severe NPDR
Severe NPDR	Any of the following
	Intraretinal haemorrhages (20 in each quadrant)
	Definite venous beading (in two quadrants)
	Intraretinal microvascular abnormalities (in 1 quadrant)
	No signs of proliferative retinopathy
PDR	Severe NPDR and one or more of the following
	Neovascularisation
	Vitreous/preretinal haemorrhage

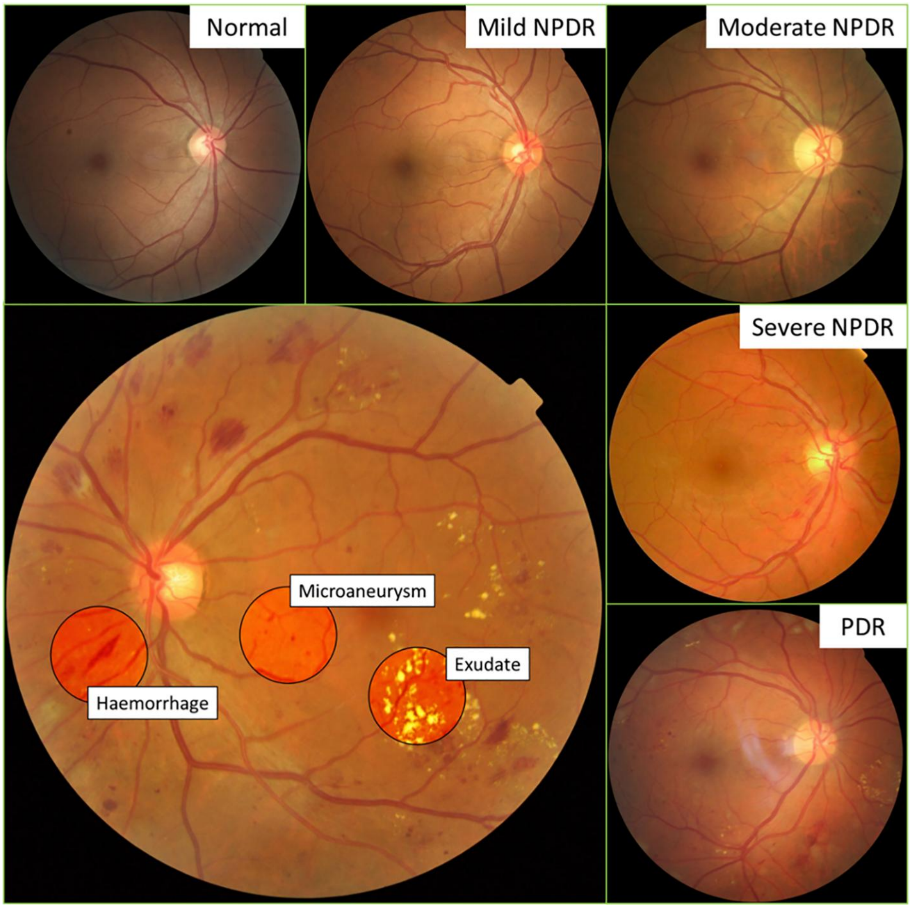

FIGURE 1 Example of images that represents the severity levels and the presence of microaneurysm, haemorrhage and exudate in a severe non-proliferative diabetic retinopathy (NPDR) fundus image

image and classify the image to a related disease category. A Deep Learning classification model that has been widely applied in medical imaging is the CNN.

The Convolutional Neural Network is a supervised network in DL that is typically used when dealing with image data. The Convolutional Neural Network has become a dominant architecture in medical image analysis published work due to the availability of huge labelled datasets and advancement of Graphical Processing Units (GPUs) leading to a significant improvement of CNN performance [12]. It comprises three types of layers which are convolutional, pooling layers, and fully connected layers. Some of the popular

CNN architectures are AlexNet [13], VGGNet [14], GoogLeNet [15], or commonly known as Inception and ResNet [16] architecture.

In CNN for DR classification, various studies have been done with different approaches from classifying retinal images based on their severity levels to classifying images based on the extracted features. More details will be highlighted in Section 3.

Accurate results in classifying medical images are important in an automated system to aid clinical care and treatment. The Convolutional Neural Network is an end-to-end solution for image classification; therefore, the model will learn the features of each class and be able to differentiate the characteristics required for each class. The strength of CNN is that the error-detected abnormalities will be propagated back to enhance the feature extraction component, resulting in improved representation [17]. It has been widely used in medical diagnosis or classification systems because the CNN is a good feature extractor, in which the network can be used to categorize medical images. It is also an economical option to save time and money on feature engineering [18], which is a process of selecting and transforming features from raw data into a form that is easier to interpret when building a predictive model.

3 | DIABETIC RETINOPATHY DETECTION AND GRADING

There have been various approaches in detecting and grading DR severity using DL which can be categorised into two categories where the first category is by training the classification model to distinguish DR grades and the second category is by using information obtained from the extracted features of lesions which are microaneurysms, exudates, and haemorrhages [19]. Diabetic Retinopathy detection determined whether the retinal image is normal, or the image is abnormal. On the other hand, DR grading classifies the image to the different levels of severity of the DR.

To classify DR grades directly, the classification model is trained to differentiate the features for each level of severity. Grading of DR is performed based on its severity level by referring to Table 1 which has been used by ophthalmologists to grade a fundus image for DR. The interpretation of the table is based on the expert in which there might be variations in opinions for each level resulting in a need of a third opinion. Diabetic Retinopathy grading to 5 levels applied by ([1, 20–42]) will be discussed further in Section 4.

However, there are occasions when researchers classify differently as they combined some levels. Some studies are more focussed on detecting the presence of DR where they classify images to 0 or 1 which refers to no DR or DR of any severity level [43] as they merged data from class (1,2,3,4) and defined them as 1. Gulshan et al. [6] focussed on training algorithms to detect referable DR which is defined as moderate and worse DR. Shankar et al. [44], Zhang et al. [45] and Hemanth et al. [46] classified the retinal images into 4 classes. However, the levels used in both studies differ from each other. Zhang et al. [45] used 4 severity levels of DR; No DR; NPDR; NPDR2PDR; and PDR while Hemanth et al. [46] classified the images into normal; macular oedema; PDR; and NPDR.

Diabetic Retinopathy detection can also be performed by detecting and segmenting lesions using information about their location such as microaneurysms, exudates, and haemorrhages as shown in Figure 1. The network is trained to learn the features of microaneurysms, exudates, and haemorrhages. Khojasteh et al. [47] classified the images into three stages by detecting the presence of exudates, microaneurysms, and haemorrhages whereas Eftekhari et al. [48] take advantage of CNN to increase the accuracy of microaneurysms detection whereby CNN works as the main classifier to extract the potential microaneurysm regions. On the other hand, Mateen et al. [49] detected exudates from DR by performing transfer learning on Inception V3, ResNet-50, and VGG-19 architectures. The collective information obtained on the lesions can be used in the future for DR classification.

Both DR detection and grading are important in a way in which both provide information of a patient suffering from diabetic retinal disease. Although no cure for DR has been developed, some treatments can be applied to prevent further damage in the eye. The doctor will determine the best possible treatment according to DR severity stages. Therefore, knowing the grade of DR for a patient will help the patient with the disease and be more aware in monitoring for any changes in the body due to diabetes.

4 | EXISTING WORKS

The general process of a CNN model in classifying fundus images to 5 severity levels are explained in Figure 2. Retinal images are split into training and testing images. Then, the images are fed into the CNN model where images are resized according to the input image layer of the CNN. A CNN model consists of convolution layer, pooling layer and fully-connected layer. There is no specific quantity for how many of each type of layer are needed in a model and fundus image will be graded accordingly. Thus, the existing works that we focussed on are the usage of DL in DR grading to 5 stages which are No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and PDR. A total of 20 papers that utilised CNN for DR grading have been reviewed. Table 2 displays the summary of studies that applies the CNN for DR grading as mentioned in Table 1 [8]. In Table 2, information regarding the dataset used, the number of images used; train-test split; model or architecture used, and their performance measures (area under the curve (AUC), accuracy, specificity, and sensitivity) were presented. The subsections will discuss in more detail the dataset, CNN architecture employed, and performance results achieved.

4.1 | Datasets

In this section, we will analyse the datasets, the number of retinal images used, and the train-test split employed by the researchers. Public and private datasets have been utilised in

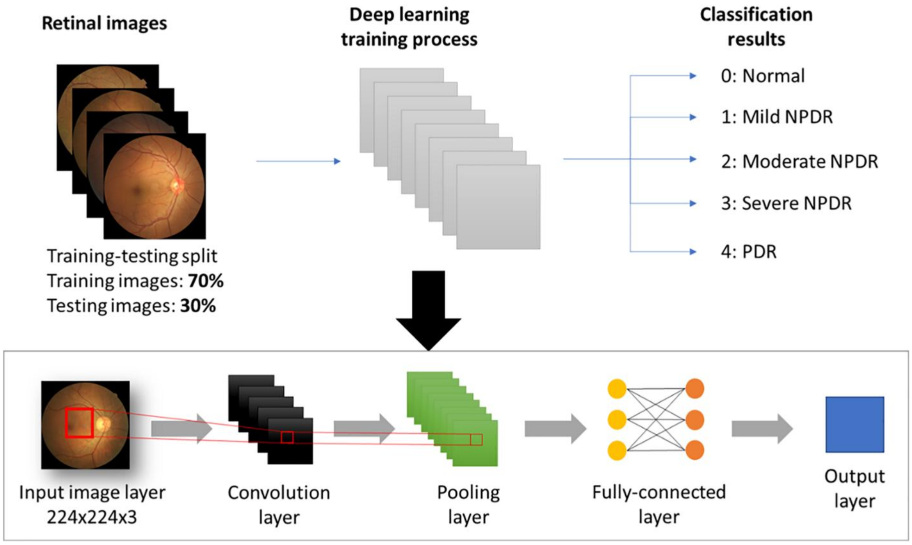

FIGURE 2 Deep Learning (DL) process in classifying images to 5 severity levels

DR grading to 5 severity levels. Public datasets that provide pre-collected retinal images for research purposes can be obtained from Kaggle, EyePACS [50], Messidor [51], Messidor-2 [52], DIARETDB1 [53], and Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 dataset.

Based on Table 1, Ref. [1, 20, 21, 23–26, 31, 32, 35, 39, 40] used the EyePACS dataset which contain 35,126 images which can be obtained from Kaggle repositories. However, Ref. [27, 28, 38], and Ref. [41] employed EyePACS datasets that contain 71,913, 37,000, 88,702 and 75,000 images. The EyePACS dataset has been graded to its severity stages labelled from zero to 4. Bhardwaj [29] used the Messidor dataset with 396 images. On the other hand, Reguant, Brunak, and Saha [1] utilised DIARETDB1 for their studies which has 89 images representing the five levels of DR grading. Messidor-2 was adopted by Ref. [32, 34, 36, 38] which is an extension of Messidor as an additional dataset with 1746 and 800 images, respectively. Ref. [35, 37, 42] utilised the APTOS 2019 dataset with 3662 images, which were used for Kaggle 2019 DR detection competition.

Ref. [22, 36], and Ref. [28] employed datasets obtained from different hospitals and institutions with 8816, 13,767, and 40,000, respectively. While Dai et al. [27] implemented three private datasets, Shanghai Integration model (SIM), China National Diabetic Compilations Study, and Nicheng Diabetes Screening Project (NDSP), for training and validation which contain 666383, 92672, and 27948 respectively.

The authors of Ref. [36, 39] split the dataset at the ratio of 90:10 for training and validation, respectively. The authors of Ref. [24, 27] split the data with the ratio of 80:20 while those of Ref. [29] divided the dataset in the ratio of 70:30. The authors of Ref. [1, 41] divided the dataset into training, validation, and testing with the ratio of 80:10:10 and 94:3:3, respectively.

In an automated system, it is important to remember that bias and prejudice must be avoided. For example, people receiving healthcare varies by race; thus, racial bias should be considered. There is biological variability that may exist for different races. A study by Li et al [54] proves that there are significant differences in several retinal parameters among Malays, Chinese, and Indians. Experiments suggest that there is a potential for bias that exists between lighter-skin and darker-skin due to average fundus pigmentation, optic disc size, and retinal arteriolar calibre [55]. The database size used varies; however, most studies operate on a relatively large dataset with more than 30,000 images. The dataset is obtained from various sources, but mostly from public sources such as Kaggle and EyePACS as they provide a large number of labelled retinal images. Before training, the ratio between training data and validation should be determined. Using a huge training dataset can provide a good result; a good ratio between the training and validation dataset is important as a good result can be obtained by having the best ratio. Although some studies did not mention the ratio that they apply in their studies, using the ratio of 90:10 can produce good results too. Ref. [36, 39] used this ratio and achieved high accuracy for their work. Therefore, with a small dataset, having the largest possible percentage of the training dataset than the validation dataset is better.

Other than that, image pre-processing is considered necessary in DL to standardise the image variation in the dataset. Some studies apply different methods to pre-process their images to enhance the quality of images. Qummar et al. [20] apply image resizing, image cropping, mean normalised image, and rotated image to their input dataset. Wan et al. [31] utilised non-local means denoising developed by Buades et al. [56] to remove noise. Studies also applied data augmentation to their dataset to increase the number of images. Data augmentation is a way of increasing the data while retaining the features in the image by performing image modifications across a sample dataset [43]. By performing data augmentation, we can increase the number of images used. To train the network for DR classification, images are enhanced to determine features for each class. Wan et al. [31] also suggest cropped images to eliminate unnecessary areas. Nneji et al. [34] applies contrast-limited adaptive histogram equalisation

TABLE 2 Summary of studies focussing on applying Convolutional Neural networks| for Diabetic Retinopathy (DR) grading

		Dataset					Performance measures			
Ref	Research focus	Database	Size	Train/ Test split	Learning	Architecture/Model	AUC	Acc	Sp	Se
Qummar et al., 2019 [20]	Encode the rich features and improve classification for different stages of DR	EyePACS	35,126	-	Ensemble	Ensemble (ResNet50, Inception-V3, Xception, Dense121, Dense169)	0.970	80.80	86.72	51.50
Wan et al., 2018 [31]	Analyse the performance of transfer learning and hyperparameter-tuning CNN models in DR classification.	EyePACS	35,126	-	Transfer learning	AlexNet	-	89.75	94.07	81.27
						VggNet-16		93.17	94.32	90.78
						VggNet-19		-	-	-
						GoogLeNet		93.36	93.45	77.66
						ResNet		90.40	95.56	88.78
						Inception-v3		93.49	93.45	96.39
Li et al., 2019 [36]	Transfer learning to identify the level of DR from retinal fundus photographs	Shanghai Zhongshan hospital and shanghai first People's hospital	8816	90:10	Transfer	Inception-v3	0.978	93.49	93.45	96.39
						VGG16	0.970	84.31	-	-
		Messidor-2	800							
Devi et al., 2021 [37]	Aggregation of deep features from multiple convolution blocks of a pre-trained model to represent retinal images	APTOS 2019	3662	-	Transfer learning	DenseNets	1.000	-	98.00	98.00
Riaz et al., 2020 [38]	CNN networks with more deep supervision to extract comprehensive feature maps	EyePACS	71,913	-	Transfer learning	DenseNets	1.000	-	98.00	98.00
		Messidor-2	1747							
Hu et al., 2019 [39]	Automatic classifier based on CNN with 2 models	EyePACS	35,126	90:10	Transfer learning	Inception ResNet V2	-	81.90	-	-
Sayres et al., 2019 [40]	Impact of DL DR algorithms on physician readers in computer-assisted settings	EyePACS	35,126	-	Transfer learning	Inception -v4	-	-	94.60	91.55
Kwasigroch et al., 2018 [41]	Diagnose DR and its current stage based on image analysis	EyePACS	37,000	94:3:3	Transfer learning	VGG-D	-	81.70	50.50	89.50
Khalifa et al., 2019 [42]	Deep transfer learning models for medical DR detection	APTOS 2019	3662	-	Transfer learning	AlexNet	-	97.90	-	-
						VGG16		97.80		
						ResNet18		97.90		
						SqueezeNet		97.80		
						VGG19		97.40		
						GoogLeNet		96.30		
Reguant et al., 2021 [1]	Discover the inherent image features and their clinical relevance	EyePACS	35,126	80:10:10	Transfer learning	Inception	-	94.00	95.00	81.00
						ResNet50		89.00	93.00	73.00

TABLE 2 (Continued)

		Dataset					Performance measures			
Ref	Research focus	Database	Size	Train/ Test split	Learning	Architecture/Model	AUC	Acc	Sp	Se
		DIARETDB1	89			Inception ResNet		94.00	96.00	83.00
						Xception		95.00	96.00	86.00
Raju et al., 2017 [21]	Detect the laterality of the eye, stage of DR for each eye and provide a diagnosis report for each eye	EyePACS	35,126	-	-	Custom CNN	-	-	92.29	80.28
Zhang et al., 2019 [22]	Developed DeepDR, a novel, and well-performing DR recognition and classification system	Sichuan academy of medical sciences and sichuan provincial peoples hospital	13,767	-	-	Ensemble	-	95.46	97.99	98.11
Ghosh et al., 2017 [23]	Uses CNN along with denoising to identify features like microaneurysms and haemorrhages	EyePACS	35,126	-	-	Custom CNN	-	85.00	-	-
Chen et al., 2019 [24]	Transfer learning to grade fundus photographs into 5 classes corresponding with 5 stages of DR	EyePACS	35,126	80:20	Transfer learning	Inception V3	-	80.00	-	-
Wang et al., 2018 [25]	Automatically differentiate the 5 stages of DR based on funduscopic images	EyePACS	35,126	-	Transfer learning	AlexNet	-	37.43	-	-
						VGG16		50.03		
						Inception V3		63.23		
Lin et al., 2018 [26]	Compare detection performance for severe DR between original fundus photographs and entropy images by deep learning	EyePACS	35,126	-	-	Custom CNN	0.920	86.10	93.81	73.24
Dai et al., 2021 [27]	DL system (DeepDR) that can detect early-to-late stages of DR	Shanghai integration model (SIM)	666,383	80:20	-	DeepDR	0.955	-	85.12	91.74
		CNDCS	92,672				0.939		82.04	90.50
		EyePACS	88,702				0.944		82.36	91.12
		NDSP	27,948				0.943		83.48	94.40
Lin et al., 2019 [28]	Framework for identifying DR based on the annotation that includes DR grades and bounding boxes of lesions	Private	40,000	-	-	Custom CNN	-	87.30	-	-
		EyePACS	75,000							
Bhardwaj et al., 2021 [29]	CNN for feature extraction and DR grading	Messidor	1200	70:30	Transfer learning	AlexNet	-	73.33	-	-
						GoogleNet		65.56		
						ResNet		65.83		
						VGG16		82.14		

(Continues)

TABLE 2 (Continued)

		Dataset					Performance measures			
Ref	**Research focus**	**Database**	**Size**	**Train/ Test split**	**Learning**	**Architecture/Model**	**AUC**	**Acc**	**Sp**	**Se**
						VGG19		80.76		
						Inception V3		87.50		
Alyoubi et al., 2021 [30]	3 CNN-based models (design and transfer learning)	-	-	80:20	Transfer learning	Custom CNN (CNN299)	-	83.30	-	-
						Custom CNN (CNN512)	0.979	88.60	97.10	88.6
						EfficientNetB0	-	82.20	-	-
Yaqoob et al., 2021 [32]	Feature map of ResNet50 that has been extracted from the pooling layer and pass to the random forest classifier	Messidor-2	1748	-		ResNet50, random forest classifier	-	96.00	-	-
		EyePACS	35,126	-	-		-	75.09	-	-
Tariq et al., 2022 [33]	Applies custom dataset to pre-trained models	Custom	-	-	Transfer learning	AlexNet	-	87.84	88.88	-
						GoogLeNet	-	91.16	92.43	-
						Inception V4	-	90.31	91.73	-
						Inception ResNet V2	-	91.61	92.53	-
						ResNeXt-50	-	97.53	97.52	-
Nneji et al., 2022 [34]	Weighted fusion deep network to automatically extract features and classify DR stages from fundus scans	Messidor-2	1200	-	-	Inception V3 (CLAHE)	-	98.50	98.00	98.90
		Kaggle	35,126	-	-	VGG-16 (CECED)	-	98.00	97.80	98.90
Majumder and Kehtarnavaz, 2021 [35]	Features are extracted by a classification model and regression model which are then concatenate and inputted to a multilayer perceptron network to classify the 5 stages of DR	APTOS	3662	-	-	DenseNet	-	82.00	-	-
		EyePACS	35,126				-	82.00	-	-

(CLAHE) and contrast-enhanced canny edge detection (CECED) to the fundus images. Contrast-limited adaptive histogram equalisation pre-processing was applied to enhance the contrast and features of a fundus image by making the anomalies more apparent while CECED enhances the edges of vessels in fundus image to reveal detailed information and characteristics [34].

4.2 | Convolutional Neural Networks architecture models

The studies mentioned either perform transfer learning or customised their own CNN for DR grading. The authors of Ref. [21, 23, 26, 28], and [30] customised their own network. The authors of Ref. [20, 22] employ ensemble learning where multiple CNN models is combined to solve DR grading. The authors of Ref. [1, 24, 25, 27, 31, 36–42], and [29] perform transfer learning on several pre-trained networks. Pre-trained CNN architectures applied for DR grading in the reviewed studies are AlexNet [13], VGGNet [14], GoogLeNet [15], Inception [57], ResNet [16], Xception [58] and DenseNet [59] architectures. This section discusses on the application of different CNN models in their works. Table 3 represents CNN model architecture used and its features.

Customised CNN designed by authors of Ref. [21, 23, 26, 28], and [30] have different characteristics. Raju et al. [21] use an input size of 448 × 448 with the filter size of 3 × 3 and 4 × 4. The model is composed of 12 convolutional layers, five pooling layers, and three fully connected layers. Ghosh et al. [23] build a model with an input size of 512 × 512. The network uses 13 convolutional layers, five max-pooling layers, three dropout layers with a mixture of 2 × 2, 3 × 3, and 7 × 7 filters. Lin et al. [26] use 4 convolutional layers with a kernel size of 5 × 5. Ayoub et al. [30] proposed 2 networks CNN299 and CNN512. CNN299 uses an input of 299 × 299 and has 4 convolutional layers and 4 max-pooling layers. CNN512 uses an input size of 512 × 512 and has six convolutional layers and six max-pooling layers. The networks differ in various ways such as the number of layers and filter size. One advantage of a custom network is its larger input size. In most datasets, the original size of images is rather large leading to resizing the images according to the input layer size. Due to the size of lesions, it may be difficult for the network to learn the feature of the tiny lesions.

Qummar et al˙ [20] and Zhang et al. [22] proposed an ensemble learning of various networks. Ensemble learning combines several bases of different networks to produce an optimum predictive model. Qummar et al˙ [20] ensembled ResNet50, Inception-V3, Xception, Dense121and Dense169 for their studies. On the other hand, Zhang et al. [22] proposed an ensemble of Inception-V3, ResNet50, Xception, Inception ResNetV2, and DenseNets by fine-tuning them before combining these networks.

Transfer learning is applied by Ref. [1, 24, 25, 27, 29, 31, 36–42] for DR grading. The authors of Ref. [1, 25, 31, 42], and [29] perform transfer learning on multiple networks and compare the performance of each model. Dai et al. [27] performs transfer learning and create DeepDR by combining three sub-networks that serve different purposes. DeepDR is built for lesion detection, lesion segmentation, and DR grading. Dai et al. [27] did not specify which networks were used but mentioned that pre-trained weights were fixed during the process. However, the rest of the work mentioned which networks were employed to give insights into how well each model performed.

Authors of Ref. [25, 29, 31, 33, 42] employ AlexNet in their works. AlexNet consists of 5 convolutional layers, three pooling layers, and 3 fully connected layers. Due to the lesser depth of the model, the model faced difficulty to learn the features for each class which can be seen in the accuracy stated in Section 4.3. As it is one of the earliest CNN architecture, studies have been amending the model to improve its performance.

Next, authors of Ref. [23, 29, 33, 37, 41 42] implement different VGG networks in their studies. VGGNet added more layers to AlexNet and used only 3 × 3 kernels to increase the model performance. Ref. [25, 29, 31, 34, 37, 41], and Ref. [42] used VGGNet-16 or known as VGG-D [14] which comprises of 13 convolutional layers and 3 fully connected layers. Ref. [29, 31], and Ref. [42] utilised VGGNet-19 or denoted as VGG-E [14]that consists of 16 convolutional layers and 3 fully connected layers. The smaller filter size makes better prediction for a smaller size of lesions in the images. The downside of a VGG network is its large size causing higher computation time.

Besides, the authors of Ref. [1, 20, 22, 32], and Ref. [42] employ ResNet in their work with the authors of Ref. [1, 20, 22, 32] specifically mentioning the usage of ResNet50 and those of Ref. [42], the usage of ResNet18. ResNet18 is composed of five convolutional layers, one average pooling layer, and a fully connected layer with a softmax layer. However, ResNet50 is a deeper network than ResNet18 with 49 convolutional layers and a fully connected layer at the end of the network.

Ref. [1, 24, 25, 29, 31, 33, 34, 36, 40, 42] employ different version of Inception architecture. Ref. [1, 29, 31, 33, 42] utilised GoogLeNet, or known as Inception presented by Szegedy et al. [15], who introduced the Inception module that allows multiple types of filter size to be used in a single image block [15]. GoogLeNet contains nine inception modules with 4 convolutional layers, 4 max-pooling layers, 3 average pooling layers, 5 fully connected layers, and 3 softmax layers. The network uses an average pooling layer with 1 × 1, 3 × 3 and 5 × 5 filter sizes and stride 3. The authors of Ref. [24, 25, 29, 34, 36] implement Inception V3 while those of Ref. [33, 40] used Inception V4. Inception V3 factorises 5 × 5 convolution to two 3 × 3 convolutions to improve computational speed. Szegedy et al. [57] incorporate Inception V2 features and factorise 7 × 7 convolutions. They also apply Batch Normalisation in the auxiliary classifiers and label smoothing to prevent overfitting. Inception V4 introduces reduction blocks to change the width and height of the grid [60].

On the other hand, Reguant et al. [1] apply Inception ResNet while Hu et al. [39] and Tariq et al. [33] use Inception

TABLE 3 CNN model architecture

Ref	Method	Model	Input size	Filter size	Features
[21]	Custom	-	448 × 448	3 × 3 and 4 × 4	12 convolutional layers, 5 pooling layers, 3 fully connected layers
[23]	Custom	-	512 × 512	2 × 2, 3 × 3 7 × 7	13 convolutional layers, 5 max-pooling layers, 3 dropout layers
[26]	Custom	-	-	5 × 5	4 convolutional layers
[28]	Custom	-	300 × 300	3 × 3	Attention fusion network
[30]	Custom	CNN299	299 × 299	-	4 convolutional layers, 4 max-pooling layers
[30]	Custom	CNN512	512 × 512	-	6 convolutional layers, 6 max-pooling layers
[20]	Ensemble learning	ResNet50, Inception-V3, Xception, Dense121, Dense169	-	-	-
[22]	Ensemble learning	-	-	-	-
[25, 29, 31, 33, 42]	Transfer learning	AlexNet	224 × 224	11 × 11, 5 × 5 and 3 × 3	5 convolutional layers, 3 pooling layers, 3 fully connected layers
[25, 29, 31, 33, 37, 41, 42]	Transfer learning	VGGNet-16	224 × 224	3 × 3 and 2 × 2	13 convolutional layers, 3 fully connected layer
[29, 31, 42]	Transfer learning	VGGNet-19	224 × 224	3 × 3 and 2 × 2	16 convolutional layers, 3 fully connected layer
[42]	Transfer learning	ResNet-18	224 × 224	3 × 3	5 convolutional layers, 1 average pooling layer, 1 fully connected
[1, 20, 22, 32]	Transfer learning	ResNet50	224 × 224	3 × 3	49 convolutional layers, 1 fully connected layer
[1, 29, 31, 33, 42]	Transfer learning	GoogLeNet	299 × 299	-	9 inception modules with 4 convolutional layers, 4 max-pooling layers, 3 average pooling layers, 5 fully connected layers, and 3 softmax layers
[57]	Transfer learning	Inception V2	299 × 299	-	Factorise 5×5 convolution to two 3×3 convolution
[24, 25, 29, 34, 36]	Transfer learning	Inception V3	299 × 299	-	Factorised 7×7 convolutions, batch normalisation in the auxiliary classifiers, and label smoothing
[33, 40]	Transfer learning	Inception V4	299 × 299	-	Introduces reduction blocks to change the width and height of the grid
[1]	Transfer learning	Inception ResNet	299 × 299	-	Inception module but incorporates residual connections of ResNet, differs in terms of hyper-parameter settings
[33, 39]	Transfer learning	Inception ResNet V2	299 × 299	-	Computational cost similar to inception v4
[1, 20, 22]	Transfer learning	Xception	299 × 299	-	Depthwise separable convolutions
[20]	Transfer learning	Dense121 and Dense169	224 × 224	-	-
[22, 35, 38]	Transfer learning	DenseNets	224 × 224	-	-
[30]	Transfer learning	EfficientNetB0	-	-	The top 2 layers are replaced with the global average pooling (GAP) layer, 2 fully connected layers, and the SoftMax layer.

ResNet V2, where the model implements the idea of the Inception module but incorporates residual connections of ResNet. The differences between these two are the hyperparameter settings and their computational cost. V1 is similar to Inception V3 for its computational cost while V2 is similar to Inception V4.

Additionally, Ref. [1, 20], and Ref. [22] employ the Xception. Xception uses Depthwise Separable Convolutions in their network which is said to be more efficient in computational time. [20, 22, 38], and [35] also apply DenseNet. The authors of Ref. [20] use Dense121 and Dense169 in their studies. Ref. [35, 38], and [22] use DenseNets. Ref. [30] also used transfer learning using EfficientNetB0 where the top two layers were removed and replaced with Global Average Pooling layer, 2 fully connected layers, and SoftMax layer.

The concept behind each model implemented differs as mentioned before; however, the goal is the same which is to achieve high accuracy in performance. Nine studies utilise the different version of Inception architectures, 6 studies use the VGG network, 5 studies employ ResNet architecture, and 2 studies use Inception Resnet, Xception, and DenseNet. Five studies built their own CNN structures. Those with relatively huge datasets such as Ref. [27] opted for custom CNN for DR grading. 12 studies use transfer learning because it reduces training time and requires fewer data to train on to increase performance compared to building their model from scratch. The existing CNN architectures used in the studies vary from one another which influences different model performance accuracy. Overall, the results obtained are similar. Convolutional Neural Network architecture such as VGGNet, AlexNet, and ResNet18 is considered as a small network; thus, they compute with a lesser time compared to other models. A more complex network includes Inception architecture, DenseNet, and Xception. These models require a longer computational time due to their complexity. From the studies, it can be deduced that certain models perform better. A more complex model such as Inception does perform better than a simple model such as AlexNet.

However, with tuning AlexNet can perform better as proven by Wan et al. [31]. Hyperparameter tuning is determining a set of optimal hyperparameters for a learning algorithm. The hyperparameter tuning technique helps to carefully select the parameter values and leads to better classification performance. Deep Learning common parameters are optimization function, learning rate, mini-batch size and number of epochs.

There are several optimization functions that can be apply before the training process begins. Some of the more commonly used are stochastic gradient descent which is applied by Ref. [5, 34], Adaptive Moment Estimation Algorithm (ADAM) [33] and RMSprop [22]. Shankar [44] mentioned that the Bayesian optimization model can be used to tune a model as it analyses the previous validation outcome in which it utilises to create a probabilistic model, which will map the hyperparameters to a probability score.

Learning rate is a hyperparameter that control changes in the model in response to the estimated error each time model weights are updated. Qummar et al. [20] uses an initial learning rate 0.01, then, it is decreased by a factor of 0.1 × 10^{-5}. Zhang et al. [22] mentions in their work that they used a learning rate of 2 × 10^{-4} while Tariq et al. [33] applies a learning rate of 1 × 10^{-5}. A learning rate that is too large can cause the model to converge too quickly, whereas a small learning rate can cause the training process to get stuck.

Mini-batch size are tuned according to where the training will be executed. It is dependent on the memory requirements of the GPU or Central Processing Unit hardware. It is usually determined by the power of 2, for example, 32, 64, 128 and so on. Zhang et al. [22] and Tariq et al. [33] use a batch size of 32. Small values led to a learning process that will converge quickly with the presence of noise in the training process. On the other hand, large values cause the process to converge slowly with accurate estimation of the error gradient. Thus, it is recommended to use the largest possible value for mini-batch size for a given a computational architecture during training to achieve the best training stability and generalisation performance.

An epoch is one cycle through the entire training dataset which decides the number of times the weights in the network is to be updated. The model used 50 epochs for training with early stopping if the model starts over-fitting [22]. However, it is noted that there is no fixed number of epochs that will improve the model performance. A smaller learning rate may require more training epochs as during each update, small changes are made to the weights. On the other hand, larger learning rates result in rapid changes and require fewer training epochs.

Hyperparameter tuning can improve the model's performance if the model is trained with optimised parameters. The hyperparameter tuning technique comes with experience as there is no right answer to which hyperparameters should be tuned and its optimum values. Overall, all the models trained have been demonstrated by achieving good results.

4.3 | Performance evaluation measurement

Performance evaluation uses measurement and analysis to answer specific questions on how well a programme is achieving its outcomes. Different metrics evaluate different characteristics of the classifier induced by the classification algorithm. A confusion matrix is a correlation between the predictions of a model and the actual class labels of data points. True positive and True negative are defined as positive and negative instances that are correctly classified. False-positive (FP) and False Negative are the number of misclassified negative and positive instances, respectively. Accuracy measures the ability of the model in identifying all samples [61]. Sensitivity is the frequency of correctly predicted positive samples among all real positive samples while specificity measures the ability of a predictor in identifying negative samples [61]. In this section, the AUC, accuracy, specificity, and sensitivity are analysed. Considering this as a medical diagnosis, a high-performing model is needed, thus, the importance of AUC, high accuracy, specificity, and sensitivity.

For custom-built CNN in Ref. [21, 23, 26, 27], and [28], the overall achieved AUC, accuracy, specificity and sensitivity achieved are higher than 0.920, 85%, 82.04% and 73.24%, respectively. The highest AUC score, specificity, and sensitivity are 0.97, 97.1%, and 88.6%, respectively, achieved by authors of Ref. [30] with their CNN512 network. Lin et al˙ [28] achieve the highest accuracy with 87.30%. Dai et al. [27] tested their network against various datasets and the best AUC and specificity achieved are 0.955% and 85.12% which is tested against SIM. When tested against NDSP, the highest sensitivity of 94.40% is achieved.

Ensemble learning was done by authors of Ref. [20, 22] in which Ref. [22] achieved the best accuracy, specificity, and sensitivity of 95.46%, 97.99%, and 98.11% while Ref. [20] achieved AUC of 0.97. By using AlexNet, Ref. [42] achieved the highest accuracy of 97.90% while Ref. [31] also achieved a specificity of 94.07% and sensitivity of 81.27%. The lowest accuracy achieved by Ref. [25] is 37.43%.

Ref. [21, 25, 26, 28, 32], and [33] utilised different versions of the VGG network. For VGG16, Ref. [42] achieves the highest with 97.80% while Ref. [25] scores the lowest accuracy of 50.03%. Wan et al. [31] obtained the highest specificity and sensitivity with 94.32% and 90.78% respectively. The lowest accuracy is obtained by Ref. [25] with 50.03% while the lowest specificity and sensitivity are by Ref. [41] with 50.50% and 89.50%, respectively. The implementation of VGG19 obtained by Ref. [24, 33] gives an accuracy of 97.40% and 80.76%, respectively.

For ResNet, the authors of Ref. [29, 31] did not mention the number of layers of ResNet used in their works. Wan et al. [31] achieved higher accuracy than Ref. [29] with 90.40%. The specificity and sensitivity obtained are 95.56% and 88.78%, respectively. Ref. [42] obtained the highest accuracy of 97.90% with ResNet18 and Ref. [29] achieved the lowest accuracy of 65.83%. The specificity and sensitivity achieve higher than 93% and 73%, respectively, overall with Ref. [31] obtaining the highest specificity of 95.56% and the highest sensitivity of 88.78%.

The authors of Ref. [1, 24, 25, 31, 36, 40, 42], and [29] employ different version of Inception architecture. Khalifa et al. [42] achieved the highest accuracy 96.30% while the authors of Ref. [1] obtained the highest specificity and sensitivity with 95.00% and 81.00%, respectively, by using GoogLeNet. Ref. [29] achieved the lowest accuracy for GoogLeNet with 65.56%. The accuracy obtained by the authors of Ref. [24, 25, 29], and [36] for implementing Inception-V3 is higher than 80% except for the authors of Ref. [25] who scores only 63.23%. The highest accuracy, specificity, and sensitivity for Inception V3 are achieved by Ref. [36] with 93.49%, 93.45%, and 96.39%, respectively. Ref. [36] also achieved an AUC of 0.978. For Inception V4, Ref. [40] obtained 94.60% and 91.55% for specificity and sensitivity, respectively. Reguant et al. [1] employed Xception and scores 95% of accuracy, 96% of specificity, and 86% of sensitivity. For SqueezeNet, Ref. [42] achieved accuracy of 97.80%. Riaz et al. [38] used DenseNets to achieve the highest sensitivity and specificity with 98% with AUC of 1.

In this work, the DL models from research studies in 2021–2022 were listed and the differences between models and its performance are compared. The highest accuracy is by Nneji et al. [34] with 98.50% when the features are extracted using Inception V3. There are 3 models that achieved a specificity of 98% which is Nneji et al. [34] and Devi et al. [36]. Nneji et al. [34] also achieved a sensitivity of 98.90% which is the highest among the models that has been reviewed in the paper. Overall, the Inception V3 model that has been trained using CLAHE images performed the best.

It was observed that most networks are high-performing models where the results achieved are good enough. The number of images and CNN models play an important role in obtaining results with high accuracy, specificity, and sensitivity. A small CNN model with a small dataset is enough to obtain good results if the tuning is done correctly. A large dataset does not guarantee high accuracy; however, increasing the dataset can improve the performance of the network. There are studies with a considerably small dataset that achieves good results due to the good quality of images. Retinal images are pre-processed to enhance the image quality and features can be easily defined. With these images, the network can learn the extracted features for each class better. Studies with a large dataset can learn variations of image features for each class despite using unprocessed images. In real-life applications, images obtained may not be high-quality images all the time; thus, having low-quality images is excellent practice to make a good performing classifying model for clinical settings.

5 | CHALLENGES AND FUTURE WORK

In this paper, 19 existing works related to DR were reviewed. All the studies presented had utilised the CNN model from DL for DR grading to 5 severity levels. The models are either custom-built CNN (Ref. [21, 23, 26, 28, 30]), ensemble learning (Ref. [20, 22]) or by performing transfer learning whereas (Ref. [1, 21, 23–29, 31, 36–42]). The customised CNN model is doable if there are enough resources, in terms of sizes, number of images as the customised CNN model or ensemble learning is done with a huge dataset. Transfer learning uses existing networks; therefore, it has proven its credibility in image classification. Transfer learning might be a better option as it does not require a large dataset and is time-saving. The challenges and future works in this area include dataset size, imbalanced data issues, CNN-based DR grading performance, interpretability of the features in the CNN model, and CNN model that is worth further exploration.

It has been noticed that there has been a rise in published studies on DR especially after DL is widely used in medical imaging [62]. Despite the different approaches used in building a DR classification system, it has been noted that huge data size is needed for the training. The current public datasets are composed of good quality images, thus, making the system to be unrealistic at the moment as retinal images in real life may not be as good as the current training images as there are more diverse populations of patients with diabetes as mentioned by Abramoff et al. [62].

Overfitting is one of the critical issues faced during training. When the number of epochs used to train a network is more than necessary, the training model learns the pattern to the point of memorising the patterns that are specific to the sample data. This causes model to be incapable of performing well on a new dataset. Due to this reason, most of the studies apply early stopping to their model [1, 9, 27]. The training is stopped as soon as the training graph starts to overfit, or the error starts to increase.

One of the limitations of applying DL in the medical field is the size of the datasets needed to train a DL model [63]. It has been observed that in public datasets, the number of images for each level is imbalanced. Mostly the number of images for No DR or normal is extremely more than the other levels which may affect the performance of the classification model. Researchers need to be careful in performing data augmentation as it can cause misinterpretation of the actual disease especially if the model is intended for real-life clinical deployment. Thus, to build a robust DL classification model, the DL algorithm needs to be adaptive to imbalance issues.

Another point that should be discussed is the ability of the model to classify images from a different dataset [34]. Despite achieving high accuracy for classification, the credibility of the model when tested against a different dataset should be questioned. The model is trained using images with differences in labelling, noise and other factors that are taken into account. For example, the accuracy of a model tested against an EyePACS is 75.09% and Messidor-2 is 96% by [41]. This indicates that the model should be sensitive towards the feature noise in a dataset [64]. Therefore, the model that has been trained by taking into account different conditions should be able to perform well, despite being taken using a low-quality fundus camera.

The performance of a CNN model for DR grading is dependent on the quantity and quality of retinal images. Creating a variety of data with different ethnicities is needed to allow the model for DR grading in clinical settings to prevent biases. The structure of the eye varies with ethnicity [65]; thus, it is needed to confirm the robustness of the model in grading retinal images for DR. Variation of data during training and validation may increase the model performance [27].

Poor image quality acquisition during mass DR screening is inevitable [66]; thus, additional image quality assessment is needed to make the model functions well in such a situation. Regardless of the image quality, DL models should be capable in classifying these types of images in order for the model to be implemented in real-life application. The DL models need to capture the features, then, differentiate and grade the images accordingly. Otherwise, images with low illumination and determining features will be missed causing misclassification. Hence, the reliability of DR grading should be questioned as mislabelled images may lead to unreliable measurement of the model's final performance.

In Ref. [67], the user does not know the features used in the CNN whereby the interpretability of the DL model is still a black-box, making the explanation of the network and processing complicated [1, 47], and [68]. The Convolutional Neural Network is only provided with the images and their associated grade without the definitions of features related to its grade. Thus, an exploration of the layers is necessary to comprehend the feature at the pooling layer of the network. This is somewhat true as a user only cares about the accuracy achieved after the training process. The training process is still a mystery as the kind of features that are learnt by the model is still unknown. It would be interesting if the features that were used for DR grading at the pooling layer were to be investigated. By doing so, the model performance could be improved by removing unnecessary layers. A detailed experiment at the pooling layer is necessary to distinguish the features used by the model in determining the grades. Thus, after the network has been trained by going into each layer, an overview of how to improve the model performance will be given. However, it is a challenge to go through individual layers; finally, it can add value to the overall operational time of the network.

Currently, the lack of a system that could accurately classify DR to 5 severity levels and detect DR lesions is a gap that needs to be filled [63]. The model's capability to be implemented in real-world clinical settings is still insufficient. Since DR screening can be performed by someone who does not know DR grading, the model should fit its purpose of assisting experts in grading retinal images. In clinical situations, the distribution of images for each severity level may be unbalanced, and the images may be of poor quality, significantly reducing the model's performance. Further experiment is necessary by using an extensive amount of training and validation data with a mixture of high- and low-quality retinal images. This can be considered as a current research challenge for researchers that needs to be investigated further.

Another major challenge in DR grading is inefficient preprocessing for preparing the training data. For accurate classification of DR images, the distinctive features need to be addressed more carefully. Due to the highly varied microvascular structure of the human retina and inefficient imaging strategy, distinguishing the target features for data classification from a massive dataset appears as a big challenge. In most cases, it is almost impossible to differentiate the background and foreground of training retinal images as the pixel intensity is nearly the same. In some cases, almost all the images have a very tiny difference in abnormal DR features which offers an ultimate challenge for grading the training samples. These issues need careful attention as a well-prepared training dataset is crucial for maximising the performance of the newly developed DR grading algorithm. To address this issue, experienced personnel and an effective pre-processing strategy need to be involved in the acquisition system.

In future work, it is necessary to explore in-depth the CNN model to enhance model performance in DR grading. The Inception model can be the focus for further exploration. Furthermore, the features used by the model to classify images will need to be studied as the number of occurrences for each class are different which will affect the learning process of the model. Quality of images plays a role in training the model in which a model should be able to work with low quality images. Hence, a CNN model with high performance and reliablity will be created for DR grading.

6 | CONCLUSION

Automated DR grading is necessary for clinical settings to assist the ophthalmologist in grading retinal images as the number of diabetic patients has been increasing. With the advancement of DL, studies on the application of CNN in DR grading have been on the rise. This article reviews existing studies that have been covered on the CNN application for DR grading. Datasets can be obtained from multiple resources, either public or private, which have already been labelled to their significant grades. Various CNN models either customised to cater to DR grading or a pre-trained model have been applied in retinal image classification to achieve high accuracy so that these models can be applied in clinical settings.

The CNN approaches in DR grading whether for the detection or grading differ in the number of classes used. Knowing the presence of DR and its progressiveness helps in deciding the best treatment for the patients. The usage of CNN models for the classification of retinal images affects the performance measures. As can be seen, each study used a different model, either customised CNN model, transfer learning, or ensemble learning. The results achieved from these models have been relatively high. However, it is noted that it is currently difficult to implement the models in clinical settings as more testing must be carried out before its usage is approved.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to University Malaysia Sarawak and Ministry of Science Technology and Innovation, Malaysia for supporting this research through Technology Development Fund 1 (TeD 1) eDana (TDF05211383).

CONFLICT OF INTEREST

It is hereby declared that there are no conflicts of interest regarding the publication of this paper.

DATA AVAILABILITY STATEMENT

Data sharing not applicable–no new data generated, or the article describes entirely theoretical research.

ORCID

Nurul Mirza Afiqah Tajudin https://orcid.org/0000-0003-4992-192X
Kuryati Kipli https://orcid.org/0000-0002-4103-0674
Muhammad Hamdi Mahmood https://orcid.org/0000-0001-8410-2052
Dayang Azra Awang Mat https://orcid.org/0000-0002-3567-6372
Rohana Sapawi https://orcid.org/0000-0003-4882-7384
Siti Kudnie Sahari https://orcid.org/0000-0002-0946-0359
Kasumawati Lias https://orcid.org/0000-0002-5534-0614
Suriati Khartini Jali https://orcid.org/0000-0003-2243-2878
Mohammed Enamul Hoque https://orcid.org/0000-0002-7369-8003

REFERENCES

1. Reguant, R., Brunak, S., Saha, S.: Understanding inherent image features in CNN-based assessment of diabetic retinopathy. Sci. Rep. 11(1), 1–12 (2021). https://doi.org/10.1038/s41598-021-89225-0
2. Choudhury, A., Gupta, D.: A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques, vol. 740. Springer Singapore (2019)
3. Kalita, J., Emilia, V.: Advances in Intelligent Systems and Computing 740 Recent Developments in Machine Learning and Data Analytics (2018)
4. Gupta, D., et al.: Computational approach to clinical diagnosis of diabetes disease: a comparative study. Multimed. Tool. Appl. 80(20), 30091–30116 (2021). https://doi.org/10.1007/s11042-020-10242-8
5. Nagaraj, P., et al.: Artificial flora algorithm-based feature selection with gradient boosted tree model for diabetes classification. Diabetes, Metab. Syndrome Obes. Targets Ther. 14, 2789–2806 (2021). https://doi.org/10.2147/DMSO.S312787
6. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, J. Am. Med. Assoc. 316(22), 2402–2410 (2016). https://doi.org/10.1001/jama.2016.17216
7. Yip, M.Y.T., et al.: Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. npj Digit. Med. 3(1), 31–34 (2020). https://doi.org/10.1038/s41746-020-0247-1
8. Wilkinson, C.P., et al.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmol. 110(9), 1677–1682 (2003). https://doi.org/10.1016/S0161-6420(03)00475-5
9. Zago, G.T., et al.: Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Comput. Biol. Med., 103537 (2019). https://doi.org/10.1016/j.compbiomed.2019.103537
10. Wang, W., Lo, A.C.Y.: Diabetic retinopathy: pathophysiology and treatments. Int. J. Mol. Sci. 19(6) (2018). https://doi.org/10.3390/ijms19061816
11. Sahiner, B., et al.: Deep learning in medical imaging and radiation therapy. Med. Phys. 46(1), e1–e36 (2019). https://doi.org/10.1002/mp.13264
12. Ker, J., et al.: Deep learning applications in medical image analysis. IEEE Access. 6, 9375–9379 (2017). https://doi.org/10.1109/ACCESS.2017.2788044
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks (2012)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 1–14 (2015)
15. Szegedy, C., et al.: Going deeper with convolutions. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12, 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
16. He, K., et al.: Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
17. Muhammad, S., et al.: Medical image analysis using convolutional neural networks A review. J. Med. Syst, 1–13 (2018)
18. Yadav, S.S., Jadhav, S.M.: Deep convolutional neural network based medical image classification for disease diagnosis. J. Big Data. 6(1) (2019). https://doi.org/10.1186/s40537-019-0276-2
19. Li, X., et al.: CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Trans. Med. Imag. 39(5), 1483–1493 (2020). https://doi.org/10.1109/TMI.2019.2951844
20. Qummar, S., et al.: A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access. 7, 150530–150539 (2019). https://doi.org/10.1109/ACCESS.2019.2947484
21. Raju, M., et al.: Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy. Stud. Health Technol. Inf. 245, 559–563 (2017). https://doi.org/10.3233/978-1-61499-830-3-559

22. Zhang, W., et al.: Automated identification and grading system of diabetic retinopathy using deep neural networks. Knowl. Base Syst. 175, 12–25 (2019). https://doi.org/10.1016/J.KNOSYS.2019.03.016
23. Ghosh, R., Ghosh, K., Maitra, S.: Automatic detection and classification of diabetic retinopathy stages using CNN. 2017 4th Int. Conf. Signal Process. Integr. Networks, SPIN. 2017, 550–554 (2017). https://doi.org/10.1109/SPIN.2017.8050011
24. Chen, H., et al.: Detection of diabetic retinopathy using deep neural network. Int. Conf. Digit. Signal Process. DSP. 2018 (2019). https://doi.org/10.1109/ICDSP.2018.8631882
25. Wang, X., et al.: Diabetic retinopathy stage classification using convolutional neural networks. Proc. - 2018 IEEE 19th Int. Conf. Inf. Reuse Integr. Data Sci. IRI. 2018, 465–471 (2018). https://doi.org/10.1109/IRI.2018.00074
26. Lin, G., et al.: Transforming Retinal Photographs to Entropy Images in Deep Learning to Improve Automated Detection for Diabetic Retinopathy, vol. 2018 (2018). [Online]. https://doi.org/10.1155/2018/2159702
27. Dai, L., et al.: A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat. Commun. 12(1) (2021). https://doi.org/10.1038/s41467-021-23458-5
28. Lin, Z., et al.: A Framework for Identifying Diabetic Retinopathy Based on Anti-noise Detection and Attention-Based Fusion, vol. 1. Springer International Publishing (2019)
29. Bhardwaj, C., Jain, S., Sood, M.: Transfer learning based robust automatic detection system for diabetic retinopathy grading. Neural Comput. Appl. 2, 13999–14019 (2021). https://doi.org/10.1007/s00521-021-06042-2
30. Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M.: Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors. 21(11), 1–22 (2021). https://doi.org/10.3390/s21113704
31. Wan, S., Liang, Y., Zhang, Y.: Deep convolutional neural networks for diabetic retinopathy detection by image classification R. Comput. Electr. Eng. 72, 274–282 (2018). https://doi.org/10.1016/j.compeleceng.2018.07.042
32. Yaqoob, M.K., et al.: Resnet based deep features and random forest classifier for diabetic retinopathy detection[†]. Sensors. 21(11), 1–14 (2021). https://doi.org/10.3390/s21113883
33. Tariq, H., et al.: Performance analysis of deep-neural-network-based automatic diagnosis of diabetic retinopathy. Sensors. 22(1), 1–15 (2022). https://doi.org/10.3390/s22010205
34. Nneji, G.U., et al.: Identification of diabetic retinopathy using weighted fusion deep learning based on dual-channel fundus scans. Diagnostics. 12(2), 540 (2022). https://doi.org/10.3390/diagnostics12020540
35. Majumder, S., Kehtarnavaz, N.: Multitasking deep learning model for detection of five stages of diabetic retinopathy. IEEE Access. 9, 123220–123230 (2021). https://doi.org/10.1109/ACCESS.2021.3109240
36. Li, F., et al.: Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. Transl. Vis. Sci. Technol. 8(6) (2019). https://doi.org/10.1167/tvst.8.6.4
37. Devi, J., et al.: Deep convolution feature aggregation: an application to diabetic retinopathy severity level prediction. Sign. Image Video Proc. 15(5), 923–930 (2021). https://doi.org/10.1007/s11760-020-01816-y
38. Riaz, H., et al.: Deep and densely connected networks for classification of diabetic retinopathy. Diagnostics. 10(1), 1–14 (2020). p24.1p. https://doi.org/10.3390/diagnostics10010024
39. Hu, W., Zhang, Y., Li, L.: Study of the application of deep convolutional neural networks (CNNs) in processing sensor data and biomedical images. Sensors. 19(16) (2019). https://doi.org/10.3390/s19163584
40. Sayres, R., et al.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmol. 126(4), 552–564 (2019). https://doi.org/10.1016/j.ophtha.2018.11.016
41. Kwasigroch, A., Jarzembinski, B., and Grochowski, M.: Deep CNN based decision support system for detection and assessing the stage of diabetic retinopathy, 111–116, (2018)
42. Khalifa, N.E.M., et al.: Deep transfer learning models for medical diabetic retinopathy detection. 27(5), 327–332 (2019). https://doi.org/10.5455/aim.2019.27.327-332
43. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. Ophthalmol. 124(7), 962–969 (2017). https://doi.org/10.1016/j.ophtha.2017.02.008
44. Shankar, K., et al.: Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. Pattern Recogn. Lett. 133, 210–216 (2020). https://doi.org/10.1016/j.patrec.2020.02.026
45. Zhang, W., et al.: Knowledge-Based Systems Automated identification and grading system of diabetic retinopathy using deep neural networks. Knowl. Base Syst. 175, 12–25 (2019). https://doi.org/10.1016/j.knosys.2019.03.016
46. Hemanth, D.J., Deperlioglu, O., Kose, U.: An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network. Neural Comput. Appl. 32(3), 707–721 (2020). https://doi.org/10.1007/s00521-018-03974-0
47. Khojasteh, P., Aliahmad, B., Kumar, D.K.: Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. BMC Ophthalmol. 18(1), 1–14 (2018). https://doi.org/10.1186/s12886-018-0954-4
48. Eftekhari, N., et al.: Microaneurysm detection in fundus images using a two - step convolutional neural network. Biomed. Eng. Online, 1–16 (2019). https://doi.org/10.1186/s12938-019-0675-9
49. Mateen, M., et al.: Exudate Detection for Diabetic Retinopathy Using Pretrained Convolutional Neural Networks, vol. 2020 (2020)
50. Cuadros, J. and Bresnick, G.: EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. 3(3), 509–516 (2009). https://doi.org/10.1177/193229680900300315
51. Ecencière, E.T.D., et al.: Feedback on a Publicly Distributed Image Database: The MESSIDOR Database, 231–234 (2014). https://doi.org/10.5566/ias.1155
52. Abràmoff, M.D., et al.: Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 131(3), 351–357 (2013). https://doi.org/10.1001/jamaophthalmol.2013.1743
53. Kauppi, T., et al.: The DIARETDB1 diabetic retinopathy database and evaluation protocol. BMVC 2007 - Proc. Br. Mach. Vis. Conf. 2007, 1–18 (2007). https://doi.org/10.5244/C.21.15
54. Li, X., et al.: Racial differences in retinal vessel geometric characteristics: a multiethnic study in healthy Asians. Invest. Ophthalmol. Vis. Sci. 54(5), 3650–3656 (2013). https://doi.org/10.1167/iovs.12-11126
55. Burlina, P., et al.: Addressing artificial intelligence bias in retinal diagnostics. Transl. Vis. Sci. Technol. 10(2), 1–13 (2021). https://doi.org/10.1167/tvst.10.2.13
56. Buades, A., Coll, B., Morel, J.-M.: Non-local means denoising. Image Process. Line. 1, 208–212 (2011). https://doi.org/10.5201/ipol.2011.bcm_nlm
57. Szegedy, C., et al.: Rethinking the Inception Architecture for Computer Vision (2014)
58. Google, C.: Xception: Deep Learning with Depthwise Separable Convolutions (2014)
59. Huang, G., Weinberger, K.Q.: Densely Connected Convolutional Networks
60. Szegedy, C., et al.: Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conf. Artif. Intell. AAAI. 2017, 4278–4284 (2017)
61. Jiao, Y., Du, P.: Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol. 4(4), 320–330 (2016). https://doi.org/10.1007/s40484-016-0081-2
62. Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal Imaging and Image Analysis.Pdf, 169–208. IEEE Rev Biomed Eng (2010). https://doi.org/10.1109/RBME.2010.2084567.Retinal
63. Alyoubi, W.L., Shalash, W.M., Abulkhair, M.F.: Informatics in Medicine Unlocked Diabetic retinopathy detection through deep learning techniques: a review. Inform. Med. Unlocked. 20, 100377 (2020). https://doi.org/10.1016/j.imu.2020.100377
64. Hazarika, B.B., Gupta, D., Borah, P.: An intuitionistic fuzzy kernel ridge regression classifier for binary classification. Appl. Soft Comput. 112, 107816 (2021). https://doi.org/10.1016/j.asoc.2021.107816

65. Bourne, R.R.A.: Ethnicity and ocular imaging. Eye. 25(3), 297–300 (2011). https://doi.org/10.1038/eye.2010.187
66. Zhou, L., et al.: Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images. IET Image Process. 12(4), 563–571 (2018). https://doi.org/10.1049/iet-ipr.2017.0636
67. Beiji, Z.O.U., et al.: Deep learning and its application in diabetic retinopathy screening. Chin. J. Electron. 29(6), 992–1000 (2020). https://doi.org/10.1049/cje.2020.09.001
68. Valizadeh, A., et al.: Presentation of a segmentation method for a diabetic retinopathy patient's fundus region detection using a convolutional neural network. Comput. Intell. Neurosci. 2021, 1–14 (2021). https://doi.org/10.1155/2021/7714351

How to cite this article: Tajudin, N.M.A., et al.: Deep learning in the grading of diabetic retinopathy: A review. IET Comput. Vis. 1–16 (2022). https://doi.org/10.1049/cvi2.12116