Computers in Biology and Medicine 135 (2021) 104599

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep learning for diabetic retinopathy detection and classification based on fundus images: A review

Nikos Tsiknakis [a,*], Dimitris Theodoropoulos [b], Georgios Manikis [a], Emmanouil Ktistakis [a,c], Ourania Boutsora [d], Alexa Berto [e], Fabio Scarpa [e,f], Alberto Scarpa [e], Dimitrios I. Fotiadis [g,h], Kostas Marias [a,b]

[a] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), 70013, Heraklion, Greece
[b] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, 71004, Heraklion, Greece
[c] Laboratory of Optics and Vision, School of Medicine, University of Crete, 71003, Heraklion, Greece
[d] General Hospital of Ioannina, 45445, Ioannina, Greece
[e] D-Eye Srl, 35131, Padova, Italy
[f] Department of Information Engineering, University of Padova, 35131, Padova, Italy
[g] Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, 45115, Ioannina, Greece
[h] Department of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, 45110, Ioannina, Greece

ARTICLE INFO

Keywords:
Artificial intelligence
Classification
Deep learning
Detection
Diabetic retinopathy
Fundus
Retina
Review
Segmentation

ABSTRACT

Diabetic Retinopathy is a retina disease caused by diabetes mellitus and it is the leading cause of blindness globally. Early detection and treatment are necessary in order to delay or avoid vision deterioration and vision loss. To that end, many artificial-intelligence-powered methods have been proposed by the research community for the detection and classification of diabetic retinopathy on fundus retina images. This review article provides a thorough analysis of the use of deep learning methods at the various steps of the diabetic retinopathy detection pipeline based on fundus images. We discuss several aspects of that pipeline, ranging from the datasets that are widely used by the research community, the preprocessing techniques employed and how these accelerate and improve the models' performance, to the development of such deep learning models for the diagnosis and grading of the disease as well as the localization of the disease's lesions. We also discuss certain models that have been applied in real clinical settings. Finally, we conclude with some important insights and provide future research directions.

1. Introduction

Diabetes Mellitus is a serious public health problem, affecting 463 million people worldwide and this number is projected to rise to 700 million by 2045 [1]. At least one third of diabetics also suffer from an eye disease which is related to diabetes, of which diabetic retinopathy (DR) is the most common one [2]. DR is characterized by progressive vascular disruptions in the retina caused by chronic hyperglycemia and can be developed by any diabetes patient, regardless of its severity [3]. It is the leading cause of blindness among working age adults around the world and it is estimated that there are approximately 93 million people with DR worldwide [4]. These numbers are expected to rise even more, mainly because of the rising prevalence of diabetes in emerging Asian countries such as India and China [5, 6].

Although diabetic retinopathy is largely asymptomatic in the early stages, neural retinal damage and clinically invisible microvascular changes progress during these early stages [7]. Thus, there is a need for regular eye screening for patients with diabetes, as timely diagnosis and subsequent management of the condition is essential [8]. Since the only preventive strategy is the control of hyperglycemia, hyperlipidemia and hypertension early detection of DR becomes even more essential [7]. In addition, regarding its treatment, currently available interventions, such as laser photocoagulation, significantly decrease the likelihood of blindness in proliferative retinopathy and diabetic maculopathy in up to 98%, if the eyes are treated at an early stage of the disease [9]. It becomes evident that the key to the delay or even prevention of blindness from diabetic retinopathy is due to early detection and appropriate treatment [10].

* Corresponding author.,
E-mail addresses: tsiknakisn@ics.forth.gr, tsiknakisn@gmail.com (N. Tsiknakis).

https://doi.org/10.1016/j.compbiomed.2021.104599
Received 7 May 2021; Received in revised form 12 June 2021; Accepted 18 June 2021
Available online 25 June 2021

Although initial diagnosis of DR may be based on functional changes in electroretinography (ERG), retinal blood flow and retinal blood vessel calibre [11], in clinical practice early diagnosis is based on fundus examination [12]. Fundus photography is a rapid, non-invasive, well-tolerated and widely available imaging technique [13] that constitutes one of the most used methods to assess the extent of DR. Utilizing fundus images, ophthalmologists observe retina lesions at high resolution in order to diagnose diabetic retinopathy and assess its severity. However, manually diagnosing DR from fundus images demands a high level of expertise and effort by a professional ophthalmologist, especially in densely populated or remote areas like in India and Africa, where the number of people with diabetes and DR is projected to increase dramatically in the next years, while the number of ophthalmologists is disproportionally low [14–17]. This has motivated the research community to develop computer-aided diagnosis systems, which will reduce the cost, time and effort needed by a medical expert to diagnose DR.

Recent advancements in Artificial Intelligence (AI) and the increase of computational resources and capabilities have created the opportunity to develop Deep Learning (DL) applications for accurate DR detection and classification. In this review article, recent DL-based methods, i.e. published after 2016, the detection and classification of DR presented and critically discussed. Although some review articles regarding the application of deep learning methods on DR have been published during the past few years [18–22], most of them focus only on specific aspects of the data analysis and modeling pipeline, as is presented in Fig. 1, which in some cases is limited to the reporting of the model's performance [21, 22], or in the commonly used preprocessing methods [19, 20], while in Ref. [22], a detailed account of the publicly available datasets is not included. These fragmented efforts call for a more detailed and integrated effort to review the technical implementations and progress in this really active research area. To this end, we present a novel holistic overview of the analysis pipeline (Fig. 1), in which apart from presenting comparative technical information regarding the development of published DL models for the classification and segmentation of fundus images, we also include a thorough analysis of the publicly available datasets, the commonly used preprocessing pipelines, as well as a presentation of models that have been applied in real clinical settings.

In particular, this article provides a thorough analysis regarding the whole analysis pipeline, starting from the data preparation and preprocessing methods presented in Section 5, followed by the data analysis stage employing deep learning models. Regarding the latter, we include two distinct sections of artificial intelligence in the context of diabetic retinopathy. The first focuses on the evaluation of strengths and weaknesses of published efforts to develop deep learning models for the classification of DR's grading in Section 6, while the second provides a similar analysis, regarding efforts to develop deep learning models for the automatic segmentation of lesions that are related to DR, such as exudates (EX), microaneurysms (MA) and haemorrhages (HE), in Section 7. As for the remaining sections of this article, Section 2 and Section 3 include introductory information regarding Diabetic Retinopathy and Deep Learning respectively. In Section 4 we provide a detailed description and assessment of the various public datasets that can be used for DL development, discussing several critical characteristics of such datasets (e.g. class balancing, grading protocol used, etc.). It is worth mentioning that the reviewed datasets target both the classification and segmentation tasks. We also provide information regarding several DL models that have been applied in a real clinical setting or have been approved by regulatory agencies for use in clinical decision processes in Section 8. Finally, we conclude this article with an elaborate discussion of our main findings and proposals for future research, in Section 9 and Section 10 respectively.

2. Diabetic retinopathy

During the early stages of diabetic retinopathy microaneurysms can be observed on the retina, and are caused by degeneration and loss of pericytes, leading to capillary wall dilatation [8, 23]. When the wall of a capillary or microaneurysm is ruptured, intraretinal haemorrhages occur. Other lesions of non-proliferative diabetic retinopathy include soft and hard exudates, intraretinal microvascular abnormalities (IRMA), venous beading and venous loops or reduplication [8, 23]. According to Stitt et al. [24], IRMAs appear as large calibre tortuous vessels in areas of ischemia and may represent attempted vascular remodelling. Finally, the distinction between non-proliferative and proliferative diabetic retinopathy is based on the presence of neovascularization, which essentially refers to the growth of new retina vessels due to ischemia to preexisting ones. Fig. 2 presents some lesions on an indicative fundus image of a retina.

At any stage of diabetic retinopathy, diabetic macular edema (DME) can occur, an endpoint which constitutes the most common cause of blindness [25]. The presence of edema is accompanied by abnormalities such as exudates within one disc diameter of the centre of the fovea, exudates within the macula, retinal thickening within one disc diameter of the centre of the fovea and microaneurysms or haemorrhages within one disc diameter of the centre of the fovea [26].

Regarding the clinical grading protocols of DR, although the gold standard is the Early Treatment Diabetic Retinopathy Study (ETDRS) grading scheme [27], its use in everyday clinical practice has not proven to be easy or practical. Several alternative scales have been proposed in an effort to improve the screening of patients and communication among caregivers [28]. The development of such simplified diabetic retinopathy severity scales in several countries [29–31], had not led to a single international severity scale so far. To that end, the Global Diabetic Retinopathy Project Group has proposed the International Clinical

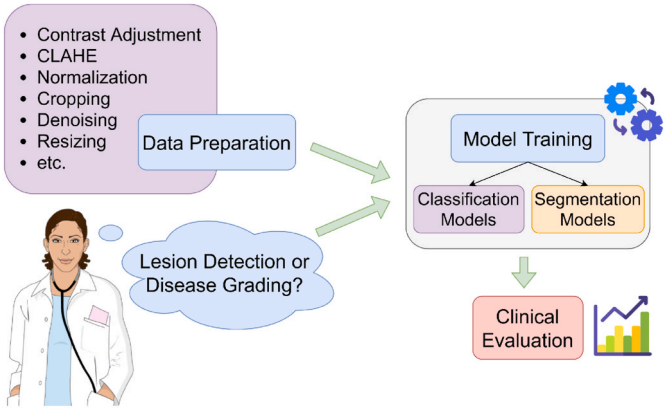

Fig. 1. Analysis pipeline of fundus images.

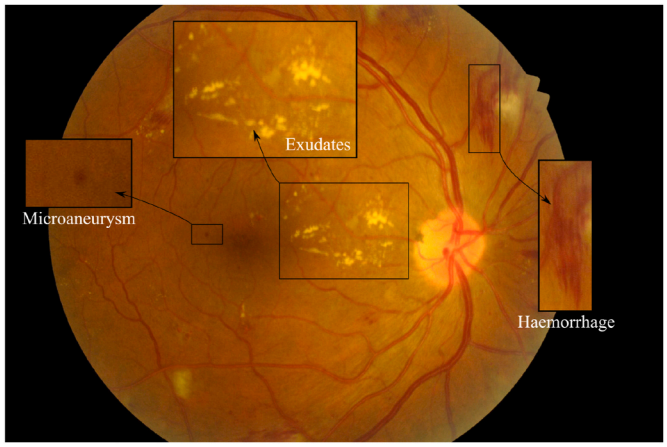

Fig. 2. Indicative DR lesions on a fundus image.

Diabetic Retinopathy Disease Severity Scale [28], which classifies DR in 5 severity scales, as shown in Table 1.

3. Deep learning

Deep Learning (DL) is a class of Artificial Intelligence (AI) methods inspired by the structure of human brain and is based on artificial neural networks. Essentially, DL refers to methods learning the mathematical representation of the latent and intrinsic relations of the data in an automatic manner. Unlike traditional machine learning methods, deep learning ones require much less human guidance, since they are not based on the generation of hand-crafted features, a task that can be very laborious and time consuming, but instead learn appropriate features directly from the data. In addition, DL methods scale much better than traditional ML methods as the amount of data increases. In this section, a short overview of some key DL concepts is provided.

3.1. Neural networks

The simplest form of a neural network refers to an Artificial Neural Network (ANN), which consists of 3 layers of neurons, one input layer, one hidden layer and a final output layer. Such networks are known as Shallow (Feed-Forward) Neural Networks, because they only have one hidden layer. In contrast, a Deep (Feed-Forward) Neural Network (DNN) consists of more than two hidden layers. Each hidden and output layer consists of several artificial neurons and every input node and hidden neuron node is connected to each neuron in the next layer through a connection link. In addition, these networks accept a one-dimensional array as their input and thus cannot be directly used with imaging data.

3.2. CNN architectures in fundus analysis

3.2.1. Traditional CNN

Convolutional Neural Networks (CNN), which unlike shallow neural networks accept 2D arrays as their input, were inspired by human vision and their concept is based on a fundamental mathematical operation, namely "convolution". The main difference of a CNN from a DNN is that for the latter all the neurons at a given layer contribute to the computation of the output of every neuron at the next layer, which is not the case for a CNN. Instead, a CNN utilizes filters or kernels to compute convolutions by sliding over a part of the original image to produce a feature map. Thus, if the size of the filter is $x \times x$, then only a window of x^2 pixels will play role in computing the value of each unit of the next layer's feature map, which directly impacts the receptive field that is defined as the region in the input space that a particular CNN's feature is affected by. Finally, the convolutional part is often referred to as "the feature extraction part" of the network, while the rest is referred to "the classification part". The former learns the imaging features that are then reduced to a one-dimensional array and fed through the latter, which essentially is a Deep Neural Network, in order to classify the input image based on the generated features.

Table 1
International clinical DR disease severity scale (ICDRDSS).

Disease Severity Level	Findings upon Dilated Ophthalmoscopy
0 = No DR	No abnormalities
1 = mild DR	Micro-aneurysms only
2 = Moderate DR	More than micro-aneurysms but less than NPDR
3 = Severe DR	Any of the following and no signs of proliferative retinopathy: 1. More than 20 intraretinal haemorrhages in each of four quadrants 2. Definite venous beading in two or more quadrants 3. Prominent IRMA in one or more quadrants
4 = Proliferative DR	One or both of the following: 1. Neovascularization 2. Vitreous/preretinal haemorrhage

IRMA = intraretinal microvascular abnormalities; NPDR = Non-Proliferative Diabetic Retinopathy; PDR = Proliferative Diabetic retinopathy Note: (1) Any patient with two or more of the characteristics of severe NPDR is considered to have very severe NPDR. (2) PDR may be classified as high-risk and non-high-risk.

3.2.2. UNet

UNet [32] architectures are more suitable for semantic segmentation than traditional CNNs, because of their ability to preserve the structural integrity of the image. In particular, they consist of a contracting path to capture the relevant context and a symmetric expanding path, enabling precise and accurate segmentation. In addition, a UNet architecture has less parameters and is faster than traditional CNNs due to the fact that it processes the image in one pass, rather than processing multiple patches in a sliding window approach, as a CNN would, and that is why such architectures are called "Fully Convolutional Networks" (FCN). Finally, it requires much less data than traditional CNNs to perform a segmentation task, which is crucial for medical image analysis, where the number of available data is much smaller than in other fields of computer vision.

3.2.3. Attention modules

It is well known that human vision and perception relies on attention mechanisms to focus on specific parts of a scene or an object instead of processing the whole scene at once [33–35]. On the other hand, traditional CNNs have yet to fully and successfully incorporate such a mechanism. To that end, many studies have recently proposed such mechanisms, called attention modules, in order to improve the performance and robustness of the models [36–39].

3.2.4. Generative Adversarial Networks

Finally, another important class of convolutional neural networks regards the Generative Adversarial Network (GAN) [40]. A traditional GAN consists of two separate models, the generative network which generates candidate samples based on the original data distribution and the discriminator which tries to distinguish the generated candidate samples from the true data distribution. Following such a training strategy, the generator is able to produce candidate samples that are closely related to the true data distribution. Application domains of GANs include image super-resolution (i.e. generate high resolution versions of the input image), creating art and image-to-image translation (e.g. transform a day image to its night equivalent) [41].

3.3. Transfer learning

Training a deep neural network is very demanding in terms of computational resources and data required. The world's largest object detection database, ImageNET [42], consists of over 14 million real life images, such as animals, devices, food, people, vehicles, etc. On the other hand, the largest dataset reviewed in this article consists of a little more than 80.000 fundus images. That difference is based on the fact that unlike images of everyday objects, medical images are very hard to obtain due to the necessary curation, annotation and legal issues involved.

Thus, training robust and accurate models can be quite difficult when it comes to medical problems. It is possible, however, to leverage models that are trained on large datasets, such as ImageNet, by transferring the obtained knowledge to another model, even if the application field differs. Transfer Learning does exactly that, i.e. improve the learning in one task by transferring knowledge from another task that is already learned [43]. Transferring knowledge from ImageNet to a medical imaging domain ultimately makes the network able to easier detect low level features of the image (i.e. edges, contours, etc.). In order to actually detect DR, one has to fine-tune (i.e. retrain) the model on the new task (i.e. new dataset), a process that, however, will be much faster and more

accurate than training it from scratch.

3.4. Ensemble learning

Another very important topic of research in AI regards Ensemble Learning, which refers to the exploitation of multiple models (i.e. base models) to produce stronger predictive results than those produced by the individual models. This learning strategy aims at reducing the generalization error of the model and is a promising technique to fuse data from multiple modalities. In ensemble learning modeling infrastructure, a diverse set of base models is operated on the same dataset, or on a subset of the given available data, towards providing single predictions from a combination of their individual outputs. There are several ensemble techniques that can be used with all the possible models, given that the task at which each individual model has been trained on is the same across all models. Such methods comprise *majority voting*, *averaging*, *bagging*, *stacking* and *boosting*.

An intuitive example of majority voting regards a classification problem, where each individual classifier produces an outcome, and the final prediction is made based on which outcome has concentrated the most votes. Averaging is used for regression problems, based on which a weighted or not average of the individual predictions are combined for the final outcome. In bagging, many models are trained each on only a subset of the original dataset, and then their outputs are combined either by majority voting, averaging or another strategy to produce the final prediction outcome. Furthermore, in stacking a separate meta-model is trained on the output predictions of the individual models, in order to produce the final prediction. Ensemble methods based on boosting aim at incorporating models that are trained multiple times based on the performance errors of previously trained and poorly performing models. Then a weighted average of the predictions is computed based on the predictive performance of the individual models.

3.5. Other learning paradigms

Other important concepts and learning strategies include Multitask Learning, Multimodal Learning and Active Learning. A simple illustration of the discussed strategies is presented in Fig. 3. By Multitask Learning one can predict the outcome for two different tasks utilizing a single data encoding network. Multimodal Learning is mostly seen in biology, pathology and radiology domains, where multiple imaging and non-imaging source (e.g. MRI, CT, molecular and clinical data) are combined for data analysis [44]. Each data modality is firstly processed by its dedicated model and then the fused features are used for training the common model. Finally, Active Learning regards the process of training the model on a small labelled subset of the data, produce the

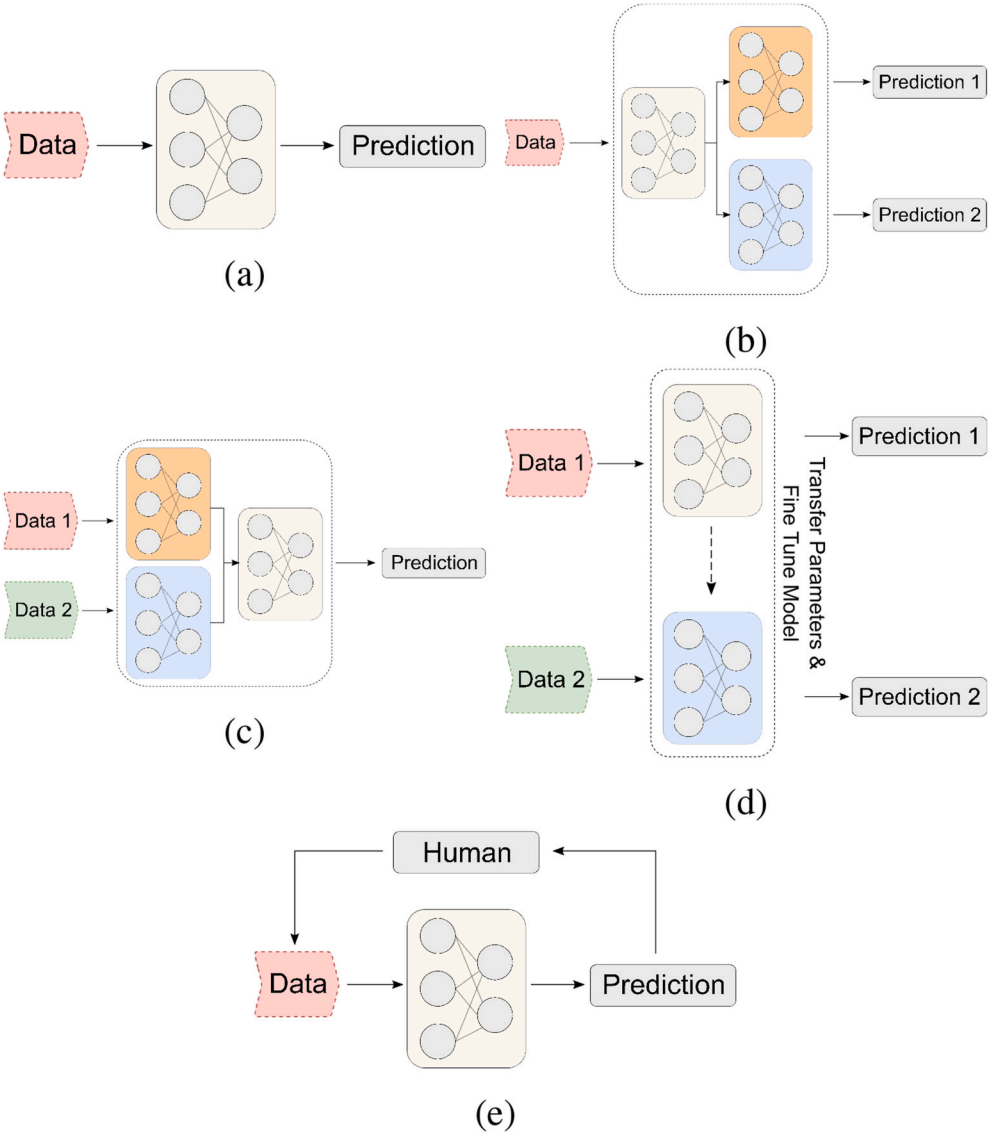

Fig. 3. Learning Strategies - (a) Simple DL model based on a single data source and predicting a single outcome, (b) Multi-task learning model based on a single data source and predicting two outcomes, (c) Multi-modal learning model based on two data sources for predicting a single outcome, (d) Transfer learning schema, in which the parameters of a source model trained on one dataset are copied to a target model for further fine-tuning purposes on another dataset, and (e) Active learning schema, where a user is queried to label new data based on a prioritization score.

predictions for the rest of the unlabelled subset of the data, prioritize them based on a given strategy and query a user for the ground truth labels of a proportion of the unlabelled data based on the prioritization score [45]. Then the model is trained on the new labelled subset of the data [45]. Active learning is mostly used in application domains where the amount of data is too large to be labelled and a priority should be given to label the data.

4. Diabetic retinopathy datasets

In this section we present various retina fundus datasets for developing and benchmarking Deep Learning Systems (DLS) in the context of diagnosing Diabetic Retinopathy. Table 2 presents details of all datasets regarding their size, resolution, Field of View (FoV) and the annotation protocol used. Table 3 presents details regarding the class-wise distribution of the severity gradings and the grading protocol used for each classification dataset. Table 4 presents similar information regarding the datasets used for segmentation purposes.

4.1. Kaggle EyePACS

Kaggle EyePACS is the most used and largest public dataset for Diabetic Retinopathy classification, containing more than 80.000 fundus images and was provided by the EyePACS platform for the Diabetic Retinopathy Detection competition which was sponsored by the California Healthcare Foundation [46]. It consists of a large number of high-resolution fundus images of the retina of both eyes, which were obtained under a variety of imaging conditions by various devices at multiple primary care sites throughout California and elsewhere. However, due to such variability, both the data (e.g. artifacts, blurring, focusing and exposure problems) and the ground truth labels exhibit noise, which was an intended goal in order to better simulate a real world scenario. The images were graded by a trained professional according to the ICDRDSS scale [28].

4.2. Kaggle APTOS 2019

Kaggle APTOS 2019 Challenge [47] dataset was collected by Aravind Eye Hospital in India's rural areas, in an effort to build powerful tools to automatically diagnose Diabetic Retinopathy and improve the hospital's ability to identify potential patients. It is the third largest dataset, consisting of 5590 images. However, one of its limitations is the large class imbalance, especially for Severe NPDR class, which has only 193 images. Just like Kaggle EyePACS dataset, due to them being collected in a real world multicentre environment, APTOS dataset also exhibits variations due to different camera settings across centres and noise both in the data (i.e. artifacts, focus problems, be under/overexposed) and the labels.

4.3. Messidor & messidor 2

The Messidor dataset [48] consists of 1200 retina fundus images which were collected by 3 ophthalmology departments, in France, between 2005 and 2006. Pupil dilation was used prior to capturing 800 images, while the rest 400 were captured without pupil dilation. Messidor 2 dataset [48, 49] contains 1058 of the images of the original Messidor dataset, as well as 690 additional images that were collected between 2009 and 2010 in the Ophthalmology department of Brest University Hospital, France.

Unlike Kaggle EyePACS dataset, the images of both datasets have very good quality, without any form of noticeable noise in them. The datasets contain an image-level medical diagnosis for each of the images, regarding the severity of Diabetic Retinopathy, but not any pixel-wise lesion segmentation information. However, their custom grading system was not consistent with the widely used ICDRS protocol, which limits its validity and applicability.

4.4. IDRiD

The IDRiD [50] dataset consists of 516 high-quality images collected at an ophthalmology clinic in Nanded, India, using a Kowa VX - 10α fundus camera. Both eyes of all subjects were dilated prior to the image capture procedure. It provides image-level grading about the severity of Diabetic Retinopathy according to the ICDRS scale and grading regarding the risk of Diabetic Macula Edema (DME) for all 516 images. It also provides pixel-wise annotations of the relevant lesions (i.e. Hard & Soft Exudates, Microaneurysms and Haemorrhages) and the optical-disc structure for 81 images of the dataset.

4.5. DDR

The DDR [51] dataset is the second largest dataset when considering the classification task, consisting of 12522 images, but it is a fairly new dataset and hasn't been used widely yet. The data were collected between 2016 and 2018 across 147 hospitals in China's 23 provinces and annotated by multiple professionals according to the ICDRDSS scale using a majority voting schema. In addition, a sixth grade was provided in order to distinguish poor quality images into a single category. However, there is a great imbalance between the healthy/moderate DR classes and the rest, i.e. mild, severe and proliferative DR, which can lead to overfitting. Regarding the relevant DR lesions, 757 images of the

Table 2
Retina fundus image datasets.

Name	Size	Resolution	Annotations	FoV	Tasks	Multiple Experts
EyePACS [46]	88702	Varying	Image Level	–	DR Grading	No
APTOS 2019 [47]	5590	Varying	Image Level	–	DR Grading	No
Messidor [48]	1200	1440 × 960 2240 × 1488 2304 × 1536	Image Level	45°	DR Grading Risk for DME	Yes
Messidor 2 [48, 49]]	1748	Varying	Image Level	45°	DR Grading	Yes
IDRiD [50]	516	4288 × 2848	Image & Pixel Level	50°	DR Grading Lesion Segmentation	Yes
DDR [51]	12.522	Varying	Image & Pixel Level	45°	DR Grading Lesion Segmentation	Yes
E-Ophtha [52]	463	Varying	Pixel Level	50°	Healthy vs Diseased Exudates and Microaneurysms Detection	Yes
DiaRetDB1 [54]	89	1500 × 1152	Pixel Level	50°	Lesion Segmentation	Yes
DRiDB [55]	50	768 × 584	Pixel Level	45°	MAS, HMs, HEs, SEs, OD and Macula Detection and Vessel Extraction	Yes
ROC [56]	100	768 × 576 1058 × 1061 1389 × 1383	Detection Level	45°	Haemorrhages and Microaneurysms Detection	Yes

Table 3
Details of the datasets used for classification.

		Classes							
Datasets	**Subsets**	**0**	**1**	**2**	**3**	**4**	**Total/set**	**Total**	**Grading Standard**
Kaggle EyePACS	Training	25810	2443	5292	873	708	35126	88702	ICDRDSS
	Testing	39533	3762	7861	1214	1206	53576		
Kaggle APTOS 2019	Training	1805	370	999	193	295	–	3662	ICDRDSS
Messidor	Whole	546	153	247	254	–	–	1200	Custom
Messidor 2	Training	–	–	–	–	–	–	1748	Custom
IDRiD	Training	134	20	136	74	49	413	516	ICDRDSS
	Testing	34	5	32	19	13	103		
DDR	Whole	6266	630	4477	236	913	–	12522 (1151 ungradable)	ICDRDSS

Table 4
Details of the datasets used for segmentation.

	Classes				
Datasets	**Healthy**	**EX**	**MA**	**HE**	**Total**
E-Ophtha	268	47	148	–	463
DiaRetDB1	5	48 Hard 36 Soft	80	54	89
IDRiD	–	81 Hard 40 Soft	81	80	81
ROC	–	–	–	–	100

dataset were annotated at a pixel-level for lesion segmentation purposes, as well as bounding boxes around them were also provided for lesion detection purposes.

4.6. E-ophtha

The E-Ophtha [52] dataset consists of 463 images, of which 268 regard healthy subjects, 148 patients with microaneurysms or other small red lesions and 47 with exudates. It has been used for automatic prediction of DR in a binary task (healthy vs diseased) [53]. However, due to the low number of images contained in the dataset compared to the larger datasets (Kaggle and Messidor), it is mostly used in the literature for developing segmentation algorithms, and not for classification ones.

4.7. DiaRetDB1

DiaRetDB1 [54] consists of 89 fundus images which were collected at Kuopio university under a controlled environment and were graded by 4 experts. However, their distribution does not reflect a typical population, since not only the data sample is small and from a single clinical site, but also all the images were captured under a controlled environment without significant variations in the capturing procedure [54].

4.8. DRiDB

DRiDB [55] consists of 50 fundus images and included annotations regarding the structure of the retina's optic disc and vessels, any present pathologies, neovascularizations and disease grading, all of which were determined by multiple experts. Although it is a fairly small dataset, it is also the most informative.

4.9. Other datasets

Additional retina fundus datasets, such as STARE [57], DRIVE [58], $ORIGA_{light}$ [59], CHASE_DB1 [60], HRF [61] and others do exist. However, they are not discussed in this paper, since their main objective regards other purposes, such as retina vessel segmentation, which are out of scope of the present review.

5. Preprocessing

As discussed in Section 4, capturing fundus images using a variety of hardware devices, under a variety of environmental conditions induces noise to the final image. In order to reduce such heterogeneity, which ultimately affects the performance of the classification model, as well as to highlight some fine details of the images, pre-processing of the images is usually a necessary step in most of the studies reviewed. In this section, we discuss several such techniques used in the reviewed literature.

5.1. Contrast enhancement

To begin with, contrast enhancement is a common preprocessing technique used for highlighting the foreground from the background, in any image processing or analysis pipeline. A simple method for contrast enhancement in fundus images is the histogram equalization [62–65], which increases the global contrast of the image, but neglects the local variations across the image. A more advanced algorithm for contrast adjustment, which takes into consideration the local variations around a specific area of each pixel, is Adaptive Histogram Equalization. However, regarding fundus imaging, Contrast Limited Adaptive Histogram Equalization (CLAHE) is more commonly used by the research community [66–69]. CLAHE solves the issue of over-amplifying the contrast in near-constant areas of the image, in contrast to the original Adaptive Histogram Equalization algorithm. By adjusting the contrast of the image, which all of the previously mentioned methods achieve, the subtle structures of the retina become more visible and easier to detect. An example of each method is illustrated in Fig. 4. Other researchers

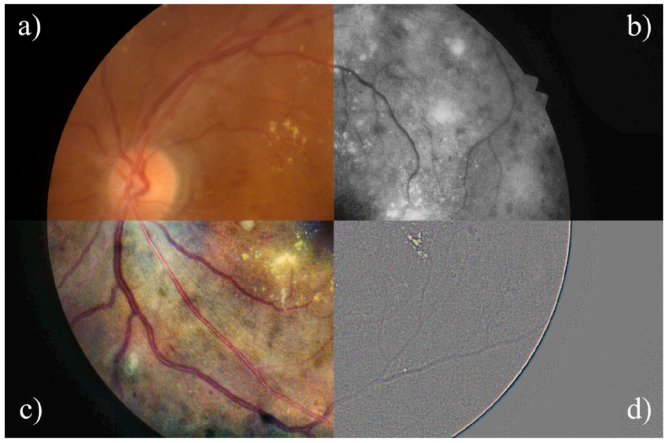

Fig. 4. A mosaic of preprocessed fundus images. Each figure illustrates the result of applying the corresponding preprocessing method to the original fundus image. a) original image, b) converting it to grayscale and applying CLAHE, c) applying CLAHE on each of the RGB channels and d) subtracting the local average color.

[70–78] subtract the local average color from each pixel and map it to 50% grayscale, to reduce differences in the lightning conditions across the images and highlight the subtle lesions, as in Fig. 4.

5.2. Denoising & normalization

Non-Local Means Denoising (NLMD) is applied by Refs. [68, 79] in order to remove potential noise in the image. However, it should be noted that although the stronger the denoising algorithm is, the more noise it will eliminate, it will also degrade the fine details of the image (i.e. the image becomes blurry).

Also, image intensity normalization is applied in order to avoid introducing bias and high training times to the network as well as to standardize the data to a particular scale (e.g. each image having a mean value of 0 and standard deviation of 1, regarding its pixels' intensity) [53, 63, 68, 72, 77, 79–86].

5.3. Color space transformation

Apart from contrast enhancement, normalization and noise reduction, transforming the color image into another color model or even simply utilizing only one of the RGB channels, have increased the model's performance. Lin et al. [81] transformed the data to entropy images, which led to the DLS outperforming the models which were trained on standard datasets. In addition, the extraction of the green channel out of the fundus color image is commonly applied, due to its rich information and high contrast in comparison with the other two color channels [64, 67, 77, 78, 87–95]. Similarly, Pao et al. [96] combined the entropy images of the grayscale image and the green channel of the original fundus image into a dual-path CNN for DR classification, effectively outperforming a CNN trained on a standard dataset.

5.4. Cropping and resizing

Furthermore, the datasets may contain images that vary in terms of resolution and aspect ratio. The images could also contain uninformative black space areas. In order to standardize the image size and to remove such black space areas, the images may be cropped, rescaled and resized to a specific resolution [51, 53, 62–64, 66, 70, 71, 73–75, 79–86, 97–101]. Bravo et al. [74] conducted their experiments using two different cropping techniques. In particular, in one experiment they crop the images so that the retina incircles the cropped image, whereas in the second experiment they cropped the largest square image inscribed in the retina.

5.5. Vessel & optic disc segmentation

Orlando et al. [102], Chudzik et al. [88] and Appan et al. [103] observed that, regarding lesion detection in a fundus image, many false-positive samples are caused by vascular branching and vessel segments misclassification. Yu et al. [92] segmented and removed the vessel structures out of the green channel of the raw image in order to enhance exudate detection, by utilizing an opening morphological operation. Chudzik et al. also utilized Otsu's thresholding [104] on the green channel of the image to segment the vessel structures, as well as morphological operations to remove noisy regions of the image. Imani et al. [89] also segmented the vessel structures, with the MCA algorithm [105], Shift Invariant Shearlet Transform (SIST) [106] and Non Subsampled Contourlet Transform (NSCT) [107] based on the structural differences in the morphology of the vessels (curved-like structures) and the exudates (spot-like structures). Morphological operations, such as opening and closing, were also used in order to refine either the vessel [102] or the lesion segmentation [89, 92]. Finally, Adem et al. [65] used the Canny Edge detector and a Circular Hough Transform method to segment the Optic Disc from retina fundus images. By removing such a complex anatomical structure, which is also similar to exudates, the performance of lesion detection improved.

5.6. Augmentation

Although DL has been proven to work well in an end-to-end manner, where the raw data is fed into a single-model pipeline, it has also been reported, that applying certain preprocessing techniques, such as those reported in this section, leads to performance improvement [66, 81, 96], especially for fundus images. In addition, due to the lack of rich and balanced datasets and in order to enhance the model's robustness and accuracy, data augmentation techniques are also used. In case of imaging datasets such techniques can refer to rotating, shifting (translation), rescaling, shearing and flipping the images, color and brightness augmentation [73, 80, 83, 84, 108], as well as the use of Generative Adversarial Networks for image synthesis [78, 109]. Regarding the reviewed articles, most utilize some augmentation method in order to increase the number of the available images and thus accelerate the training of the model.

6. Diabetic retinopathy classification

The main objective of DR classification is on the detection of diabetic retinopathy and its grading using fundus images of the retina. As described in Section 1, DR is graded by physicians with respect to a 5-class protocol. In this section we present information regarding published deep learning models that detect and grade the disease's severity on the image level.

6.1. Grading scale

In Section 2 we presented the common 5-class grading scale that ophthalmologists utilize in order to grade a fundus image regarding the DR disease. However, there are cases in which researchers decided to classify them differently, by ultimately merging several classes together. In particular, authors in Ref. [66] conducted experiments regarding DR grading with respect to a 2-class (i.e. detecting the presence of the disease), 3-class (i.e. no DR, mild DR and severe DR) and a 4-class (i.e. no DR, mild DR, moderate DR and severe DR) classification. They defined referable (i.e. the patient should be referred to an ophthalmologist) DR when at least moderate NPDR lesions are observed and vision threatening DR when at least severe NPDR or PDR lesions are observed. In Ref. [62], the authors created a 4-class scale in order to encode similar clinical manifestation between the different stages of the traditional scaling protocol (ICDRS), among other reasons. Islam et al. [71] developed two binary classification models, one for detecting the presence of the disease (healthy vs diseased) and one for grading its severity (grade 0,1 vs 2,3,4). Li et al. [51] used an additional class (6 in total) to classify ungradable images as well. Table 5 summarizes the different classification strategies utilized in the reviewed literature grouped with the corresponding references. Binary classification (i.e. detection of referable DR) is commonly formulated as classifying grades '0,1 vs 2,3,4′ or grades '0 vs 1,2,3,4'. Most reviewed articles utilize the '0,1 vs 2,3,4′ grading schema regarding the binary classification task, due to the 'mild DR' misclassification problems, which are discussed further below in

Table 5
Classification strategies.

Classification Strategy	Reference
2-class	[53, 62, 66, 70, 71, 73, 75, 79] [85, 86, 97, 98, 100, 110, 111]
3-class	[66, 75, 112]
4-class	[62, 66, 108]
5-class	[63, 64, 72, 74, 79–84] [86, 97–99, 101, 113–117]
6-class	[51]

Section 6.3.5.

6.2. Performance evaluation metrics

The performance of a binary classification model can be represented by the confusion matrix [118]. Each measure in the confusion matrix is calculated based on the predictions and the ground truth. Based on these measures, some more specialized metrics are defined, such as Sensitivity, Specificity, Accuracy, Precision, F1_score and Cohen's Kappa. In addition, the Receiver Operating Characteristic (ROC) curve presents the performance of a binary classifier by plotting its Sensitivity against its Specificity at various thresholds settings regrading the classification outcome (i.e. at which probability a given sample is considered as a positive or negative outcome). Finally, Area Under the Curve (AUC) measures the area underneath the entire ROC curve, providing an aggregate performance measure across all classification thresholds.

6.3. Approaches for model development

In this section, the various DL methods for DR classification are presented. Table 6 presents the well-established architectures (either using their pre-trained versions or not) that were used as the backbone of each proposed classification model. The most utilized architectures are VGG16, the Inception family and ResNet. We urge the reader to read an in-depth analysis of the popular architectures for deep learning based image analysis, which was recently published by Alzubaidi et al. [119].

6.3.1. Generic DL approaches

Pratt et al. [82] published one of the first studies employing a CNN based model for the quinary classification of DR (5-class – similarly to the clinical grading protocol). The authors used a class-weighted strategy to update the parameters during backpropagation, for every batch, in order to compensate for the class imbalance in the dataset and reduce over-fitting.

Islam et al. [71] converted the quinary (5-class) classification problem to a regression problem, in order to better predict NPDR and PDR cases. In addition, the authors developed a blending network, by combining the feature vectors of the CNN for each eye, in an attempt to improve the performance of the model. Similarly, Torre et al. [101] developed a CNN model, which analyzed the images of both eyes and effectively combined those representations, in order to perform the classification. They also proposed using small convolutions, and adaptations to the network's architecture in order to have a final receptive field as similar as possible to the original image's size. Raju et al. [84] also reported that when using smaller (4x4) filters in the Conv2D layers, the DR classification performance was better, due to the smaller sized lesions, such as microaneurysms. Inception modules have also been utilized in Refs. [66, 70, 110] in order to extract features at different resolutions, in an attempt to capture relevant lesion marks, which in turn vary in size.

Gulshan et al. [110] utilized a pre-trained InceptionV3 model for DR and DME classification on a dataset consisting of 128.175 images, which were retrospectively obtained from EyePACS in the US and 3 eye hospitals in India, and of which 33.246 were referable, while they tested their model on two external datasets. The images were graded between three and seven times by a pool of 54 US licensed ophthalmologists and ophthalmology senior residents for the quality of image and the presence and severity of DR and DME. Firstly, the effects of training the model on subsets of dataset with varying size were examined, with the performance reaching a plateau with the training size at around 60.000 images (with 17.000 referable). Secondly, a second subsampling experiment was performed, regarding the existence of multiple ground truth grades per image, which indicated that the performance benefited from a majority voting on those multiple grades per image. On the other hand, Krause et al. [115] determined how the use of an adjudication grading system for the development of the ground-truth labelling affects the algorithm's training performance. They used the pre-trained model by Gulshan et al. [110], fine-tuning it on a small dataset, on which an adjudication grading protocol was applied. The authors reported that even a small set of adjudicated images, allowed a slight performance improvement when using adjudication as the ground truth development standard compared to majority voting.

Table 6
Established CNN architectures used in literature.

Architectures	Reference
ResNet	[51, 62, 63, 72, 97, 120]
AlexNet	[63, 66, 99]
VGG16	[51, 63, 66, 98, 99, 117]
GoogLeNet	[51, 63, 66]
InceptionV3	[62, 70, 72, 99, 110]
InceptionV4	[74, 80, 117]
Xception	[62, 72]
Inception-ResNet	[62, 116]
DenseNet	[51, 62, 72]
SE-BN-Inception	[51]

Attention modules have also been utilized for improving the detection performance of CNNs. Zhao et al. [63] utilized an attention mechanism and a bilinear strategy, in order to train a CNN and improve the classification performance on subtle regions. Wang et al. [116] also utilized an attention mechanism to generate attention maps, which were then used by a Crop-Network, which zoomed in the highest attention regions to further improve the classification accuracy. Li et al. [85] proposed a novel architecture, which focused on jointly detecting DR and DMR, by utilizing attention modules to explore inter-disease correlations. Lin et al. [120] developed a deep learning pipeline for lesion detection, which is then used in par with the original fundus image for DR severity classification. During the detection phase, a lesion clustering method was used in order to decrease the impact of missing lesion annotations. The lesions maps from the detection model are fused with the feature maps of the classification model by the Attention Fusion Network, which also weights the importance of each lesion area. Finally, Zhou et al. [121] proposed a collaborative weakly-supervised learning model to improve the performance of DR's grading and lesion segmentation with an attention mechanism from image-level annotated data.

Furthermore, one would want to train the model with very high resolution fundus images, in order for small lesions to be easier detected. However, the computational complexity as well as the vanishing/exploding gradient problem of deep CNNs forbid this. On the other hand, directly downsampling the images leads to a huge information loss. Zhou et al. [113] developed a novel architecture, Multi-Cell Multi-Task CNN (M^2 CNN), consisting of an Inception-ResNet-V2 stage connected to a Multi-Cell stage, which gradually increases the depth and the kernel size of the network along with the input image's resolution, in order to capture high-resolution details. Finally, a Multi-Task stage is applied, during which both a classification and a regression score are computed. Doing so, the authors formulated a more appropriate training loss function, based on the consideration that DR is a gradually progressing disease and thus discrete labelling can be misleading.

Li et al. [114] experimented with several image resolutions, concluding that the performance of the model increases logarithmically with respect to higher input image resolutions. However, as the input image resolution increases, the complexity of the network also increases exponentially. Thus, given the complexity constraints, the optimal image's resolution was 896 × 896, which boosted the performance of the algorithm especially for the correct classification of the mild DR case, which depends on extracting subtle features.

6.3.2. Transfer learning approaches

A major issue regarding deep learning and especially DL applied on medical imaging regards the availability of sufficient data to train the models. As discussed in Section 3, it is possible to overcome this issue by

transferring knowledge from one field where there is plenty of data (i.e. computer vision), to another with limited data (i.e. medical imaging). Many of the reviewed articles have utilized a transfer learning approach to develop their classification models [51, 62, 66, 72, 74, 75, 80, 83, 86, 99, 100, 110, 117].

Wan et al. [83] compared several pre-trained models, which were fine tuned on the Kaggle EyePACS dataset. They reported the best results for VGGNet-s architecture, achieving an accuracy of 95.68%, specificity of 97.43%, sensitivity of 86.47% and an AUC of 0.979, outperforming other more complex architectures. Hagos et al. [70] used transfer learning by utilizing an ImageNet pre-trained InceptionV3 model. The authors fine-tuned the classifier on a small balanced subset of Kaggle EyePACS dataset. Wang et al. [99] also reported that the InceptionV3 model achieved the best results (63% Accuracy), when transfer learning was applied on a dataset of 166 images of Kaggle EyePACS dataset. Others [51, 66], have also reported the best results when using a pre-trained Inception backbone network (GoogLeNet, InceptionV3, InceptionV4).

6.4. Ensemble learning approaches

Ensemble learning has also played an important role in developing robust and powerful AI frameworks for DR classification, by combining the advantages of several classifiers [53, 63, 70, 72, 73, 80, 100]. Ensemble learning has been reported to perform better than the respective standalone models due to the information gain caused by their complementarity. That indicates that the different base models can implicitly learn different levels of semantic representations, either because of the differences in their architecture as in Refs. [62, 100], or the training procedure as in Ref. [73].

Zhang et al. [62] developed two ensemble models, one for the identification of the disease (binary classification) and one for the grading of the disease (quinary classification). The individual models were based on several pre-trained networks, which acted as the feature extraction part, and a custom standard dense neural network, which acted as the classifier. The ensemble models outperformed the individual ones in both tasks, achieving a sensitivity of 98.10% and specificity of 98.56%. The authors also note that, the 'stronger' the base learner was (pre-trained network), the higher the performance was, in general. In addition, a dual ensemble (ensemble of the ensembles) performed better than a single ensemble in some cases.

Jiang et al. [100] developed an ensemble model, using the Adaboost classifier on 3 models, which were based on the InceptionV3, ResNet152 and Inception-Resnet-V2 architectures. they trained the model on a private dataset, which was developed in collaboration with Beijing Tongren Eye Centre. The ensemble model outperformed the individual models, achieving a Sensitivity = 85.57%, Specificity = 90.85%, Accuracy = 88.21% and AUC = 0.946. However, InceptionV3 performed better in terms of Specificity, which was 91.46%.

Quellec et al. [73] trained a CNN model, which was exported at multiple iterations during the training procedure, because as the authors claim, each unique lesion type is optimally detected at different training iterations. Subsequently, they combined the saved models using ensemble learning (Random Forest Classifier) to predict DR's severity score.

6.4.1. Interpretable DL approaches

Although deep learning has been proven to be very effective and accurate in analyzing medical images, even surpassing human performance in some tasks [110], its clinical use has yet to be widely accepted. The main reason for that regards the fact that deep learning models produce a prediction without explaining the reasoning behind it, which is a crucial step in gaining the clinician's trust. Regarding the scope of this review article, there have been several attempts to develop an interpretable model for predicting DR.

Quellec et al. [73] utilized a modification of the sensitivity criterion [122] in order to produce a heatmap, visualizing the contribution that each pixel had in the prediction of the output. Moreover, by analyzing the classification results and the heatmaps produced, the authors report that the performance trajectory experiences some leaps roughly at the time that it learns to identify the different lesion types. It should be noted that the most obvious lesions are detected earlier in the training loop, while the more subtle lesions later.

Jiang et al. [100] build an interpretable ensemble classifier, by utilizing the Class Activate Maps (CAMs) [123] technique for each individual model as well as for the ensemble model. Similarly, Torre et al. [101] introduced a receptive field score distribution model, which scores the importance of each pixel of the input image in the final classification prediction.

Sayres et al. [117] evaluated the performance of ten ophthalmologists, under 3 conditions: (a) the physicians were provided with the raw fundus images, (b) the grading results of the DLS were also made available to them and (c) the grading results and an interpretable heatmap were also provided. The heatmaps regarded the pixel-wise contribution to the final prediction, which in turn indicates possible lesions. They measured three primary outcomes: diagnosis accuracy, subjective confidence in DR grading and time spent grading. They found a trend toward higher accuracy and confidence, but also higher grading times, with model assistance. As readers gained more familiarity with model assistance, there was a trend toward increased accuracy and decreased grading time. An increased sensitivity was also observed without a significant impact on specificity. Across all images, their results indicated that the grades-only condition provided a stronger benefit than grades plus heatmap.

6.4.2. Mild DR misclassification

Lam et al. [66] developed several classification models regarding binary, ternary and quinary labelling tasks. Although they achieved high performance, with respect to sensitivity and specificity, regarding the 'no DR′ or 'severe DR′ cases, they achieved only 7% sensitivity for the 'mild DR′ case. They managed to increase that percentage to almost 30% by preprocessing the images, with the cost of dropping the performance for the other 2 classes by an amount of 10%. Others also noted that the misclassification was more common for mild DR than the other classes [51, 53, 63, 72, 79, 82], which confirms that mild DR detection is a very challenging problem and the intricate details of the disease are harder identified, because their size and number are very small (1% of image) [53].

6.5. Evaluation performance of published models

Table 7 summarizes the performance results at the highest sensitivity point of the ROC curve of the classification methods as they were reported in the corresponding papers.

7. Diabetic retinopathy lesions segmentation & detection

While classification models are essential in detecting and grading DR, which effectively accelerates DR screening, detecting and segmenting relevant lesions at the pixel level is also a crucial stage of a screening pipeline. Identifying such diseased areas on the retina plays a significant role in diagnosing and treating diabetic retinopathy, as those are the main findings an expert ophthalmologist observes, as discussed in the introductory section. Thus, in this section we present information regarding published deep learning methods focused on the automatic segmentation of lesions that are related to DR, such as exudates, microaneurysms and haemorrhages.

7.1. Performance evaluation metrics

When considering a segmentation problem, the ground truth information relates to every pixel of the image instead of the entire image, as

Table 7
Performance of the reviewed classification models.

Reference	Classification (Dataset)	Best Architecture	Accuracy	Sensitivity	Specificity	Precision	AUC	F1	Kappa	QWK
[66]	2-class (EyePACS)	GoogLeNet	–	95%	96%	–	–	–	–	–
	5-class (EyePACS)		–	29%–85%	–	–	–	–	–	–
[70]	2-class (EyePACS)	Inception V3	90.9%	–	–	–	–	–	–	–
[124]	2-class Referable DR (Private)	Ensemble	–	92.2%	92.5%	–	0.97	–	–	–
	2-class Vision Threatening DR (Private)		–	96.2%	98.9%	–	0.987	–	–	–
[53]	2-class (EyePACS)	Custom CNN & Decision Tree	–	94%	98%	–	0.97	–	–	–
	2-class (Messidor2)		–	90%	87%	–	0.94	–	–	–
	2-class (E-Ophtha)		–	90%	94%	–	0.95	–	–	–
[79]	2-class (EyePACS)	Custom CNN	95%	–	–	–	–	–	–	–
	5-class (EyePACS)		85%	39.5%–95%	–	36.5%–88.2%	–	–	0.754	–
[82]	5-class (EyePACS)	Custom CNN	75%	30%	95%	–	–	–	–	–
[111]	2-class (EyePACS)	Custom CNN	94.5%	–	–	–	–	–	–	–
[83]	5-class (EyePACS)	VGGNet	95.68%	86.47%	97.43%	–	0.979	–	–	–
[71]	2 Diseased (EyePACS)	Custom CNN	–	94.5%	90.2%	–	–	–	–	–
	2 Risk (EyePACS)		–	98%	94%	–	–	–	–	–
	5-class (EyePACS)		–	–	–	–	0.844	0.743	–	0.851
[97]	2-class (FINDeRS)	Custom CNN	95.71%	76.92%	100%	–	–	–	–	–
	3-class (FINDeRS)		60.28%	65.40%	73.37%	–	–	–	–	–
[72]	5-class (EyePACS)	Inception V3	–	80%	–	76%	–	0.77	–	0.64
[98]	2-class (EyePACS)	Modified VGGNet	–	92%	72%	–	0.923	–	–	–
	2-class (Messidor)		–	99%	71%	–	0.967	–	–	–
[99]	5-class (EyePACS)	Inception V3	63.23%	–	–	–	–	–	–	–
[73]	2-class Referable DR (EyePACS)	Custom CNN	–	–	–	–	0.954	–	–	–
[100]	2-class (Private)	Ensemble	88.21%	85.57%	90.85%	–	0.946	–	–	–
[51]	6-class (DDR)	ResNet	4%–95.74%	–	–	–	–	–	0.65	–
[108]	4-class (Messidor-2)	IDx-DR	–	96.8%	87%	–	0.98	–	–	–
[117]	5-class (Custom EyePACS)	Inception V4	–	91.55%	94.69%	–	–	–	–	–
[112]	2-class Referable (Private)	Modified VGGNet	–	90.5%	91.6%	–	0.936	–	–	–
	2-class Vision Threatening (Private)		–	100%	91.1%	–	0.958	–	–	–
[101]	2-class (Messidor-2)	Custom CNN	91%	91.1%	90.8%	88.4%	–	0.896	–	–
	5-class (Messidor-2)		–	–	–	–	–	–	–	0.832
[110]	2-class (EyePACS)	Inception V3	–	97.5%	93.4%	–	–	0.991	–	–
	2-class (Messidor-2)		–	96.1%	93.9%	–	–	0.990	–	–
[115]	2-class Moderate or Worse (EyePACS)	Inception V3	–	97.1%	92.3%	–	0.986	–	–	–

(continued on next page)

Table 7 (*continued*)

Reference	Classification (Dataset)	Best Architecture	Accuracy	Sensitivity	Specificity	Precision	AUC	F1	Kappa	QWK
	2-classMild or worse (EyePACS)		–	97%	91.7%	–	0.986	–	–	–
	5-class (EyePACS)		–	–	–	–	–	–	0.84	–
[63]	5-class (EyePACS)	Custom ResNet CNN	–	–	–	–	–	0.5436	–	–
[116]	5-class (EyePACS)	Custom CNN	–	–	–	–	–	–	0.865	–
	5-class (Messidor)		–	–	–	–	–	–	0.854	–
[84]	5-class (EyePACS)	Custom CNN	–	80.28%	92.29%	–	–	–	–	–
[85]	2-class (Messidor)	Custom CNN	92.6%	92%	–	90.6%	0.963	0.912	–	–
[75]	2-class (EyePACS)	VGG16	83.68%	54.47%	93.65%	–	–	–	–	–
[64]	5-class (EyePACS)	Custom CNN	–	–	–	–	–	–	–	0.3996
[86]	2-class (Private)	Inception V3	–	2/2	92%	12%	–	–	–	–
[81]	5-class (EyePACS)	Custom CNN	86.10%	73.24%	93.81%	–	0.92	–	–	–
[80]	5-class (EyePACS)	Ensemble	80.8%	51.5%	86.72%	63.85%	0.97	0.5374	–	–
[120]	5-class (Private)	Custom CNN–	87.3%	–	–	–	–	–	0.875	–
	5-class (EyePACS)		–	–	–	–	–	–	0.859	–
[121]	5-class (IDRiD)	Custom CNN	91.3%	–	–	–	–	–	0.905	–
	5-class (EyePACS)		89.1%	–	–	–	–	–	0.872	–
[102]	Healthy vs Diseased (Messidor)	Ensemble	–	89.3%	–	–	0.911	–	–	–
	Referable DR (Messidor)		–	93.5%	–	–	0.972	–	–	–
[125]	Referable DR (Private)	EyeArt	–	95.7%	54.5%	–	–	–	–	–
[62]	Healthy vs Diseased (Private)	Ensemble	97.67%	97.64%	98%	97.6%	0.9862	0.9769	0.953	–
	5-class (Private)		95.46%	98.11%	97.99%	95.29%	–	0.9542	0.9318	–
[74]	Healthy vs Diseased (EyePACS)	Inception V4	72.41%	–	–	–	–	–	–	–
	Referable DR (EyePACS)		86.25%	–	–	–	–	–	–	–
	Healthy vs Mild DR (EyePACS)		62.86%	–	–	–	–	–	–	–
	Healthy vs Mild vs Higher (EyePACS)		72.18%	–	–	–	–	–	–	–
	Moderate vs Severe vs Proliferative (EyePACS)		69.75%	–	–	–	–	–	–	–
	5-class (EyePACS)		45%	–	–	–	–	–	–	–

in a classification task. Because in most cases the background of the image (i.e. healthy part of retina) prevails the foreground (i.e. actual lesions), utilizing the traditional metrics at the pixel-level, i.e. accuracy, sensitivity, specificity, etc., would be misleading. This is due to the fact that since most of the ground truth image refers to the healthy part of the retina and only a small proportion of the pixels refer to lesions. Hence, the pixel-wise accuracy of a segmentation algorithm would continuously be almost perfect, without necessarily correctly detecting the relevant lesions, just because the background is mostly matched with it self. Thus, the following metrics are most suitable to evaluate the performance of a segmentation model, instead of the traditional classification-oriented ones.

A metric that is commonly used in segmentation problems is the Intersection-over-Union (IoU), which is computed by dividing the overlapping area by the area of union between the predicted (P) and ground truth (G) segmentation areas, $IoU = \frac{P \cap G}{P \cup G}$. Its value ranges from 0 to 1, with 1 signifying perfect match, and 0 meaning completely disjoint. The evaluation metric is then calculated by averaging the *IoU*s of every class. Another metric is the DICE coefficient, which is computed by dividing the double overlapping area between P and G by the total number of pixels in both areas, $DICE = \frac{2*|P \cap G|}{|P|+|G|}$. Its value is also equal to the value of the $\boldsymbol{F_1}$**score** metric. The DICE coefficient is similar to the IoU, also ranging from 0 to 1. As a matter of fact, DICE and IoU are positively correlated, meaning that although their value may not be the same, they will indicate towards the same result. What this means is that when one metric indicates that a *classifier A* is better than another *classifier B*, the other metric does it too. However, their difference emerges when quantifying how much better is one classifier from another. Based on these metrics, one can calculate the lesion-level detection accuracy, sensitivity, specificity and other classification-related metrics of a segmentation model by setting a threshold for each of the IoU and DICE metrics, above or below which a true positive or a false negative detected lesion is counted.

Free-response Receiver Operating Characteristic (FROC) curve is a graphical representation of the model's performance at all decision thresholds. It is similar to ROC curves, but in FROC's case a threshold definition is needed regarding a detected region to be considered as true or false positive/negative. For example, one could set a 50% overlap between the annotated and detected regions to indicate a true positive.

In addition, there are also other metrics for evaluating a segmentation method, such as the Hausorff Distance (HD) [126] and its variants, i.e. Average Hausorff Distanec (AHD) and Hausdorff quantile method, the Euclidean Distance Difference between the centre of masses of the two segmentation masks (ΔCMD), the Surface Distance, etc. However these metrics will not be covered in depth in this review, because none of them are mentioned or used in the reviewed papers. We urge the reader to study the following articles for more in-depth information of these metrics [127–129].

It should be noted that many of the reviewed studies report their evaluation results using metrics that are most suitable for a classification task, as reported in Section 6.2. Although we discussed the problems of such metrics in a segmentation task at the beginning of this section, we included those studies in our review, but we recommend caution when interpreting their results.

7.2. Approaches for model development

7.2.1. Generic DL approaches

Xue et al. [130] proposed a deep membrane system for multitask segmentation of microaneurysms and exudates. In particular, they utilized Mask R-CNN [131] for implementing each new hybrid membrane structure, which were a combination of tissue-like [132] and cell-like [133] membrane systems. Guo et al. [134] proposed LWENet, a lightweight segmentation network utilizing an encoder-decoder architecture and having 10 times fewer parameters than other popular architectures, i.e. DeepLab v3+ [135], FCRN [136] and HED [137]. At the same time, pre-training the encoder on the DDR classification dataset, helped increase the $F1_{score}$ at the segmentation task by almost 10%.

Guo et al. [138] also proposed L-Seg network in order to simultaneously segment all four DR related lesions in a fundus image, i.e. Soft/Hard Exudates, Haemorrhages and Microaneurysms. Their model's output consists of 4 individual segmentation maps, one for each lesion type. They also propose a multi-channel bin loss function that combines

all four outputs, to avoid class- and loss-imbalance issues. They utilized several feature maps of the network to incorporate multi-scale analysis and handle lesions of different sizes, as well as a weighted fusion module to integrate all this information and effectively analyze complex lesions. In order to properly up-sample each feature map, a deconvolution layer is used, in par with a hidden loss function to avoid gradient vanishing, which is known as deep supervision. Although they report very good results in terms of AUC in contrast to other competing networks, they also state that there is a serious misclassification problem in small lesions, such as microaneurysms, etc.

Otalora et al. [139] proposed a model that was based on Active Learning to train a CNN for exudate detection in loosely labelled fundus images. In particular, the images were cropped into several smaller patches, of which only a few have relevant ground-truth information. The rest unlabelled patches are ranked based on their "interestingness", as the authors mention, which essentially encodes how much information the patch has regarding the specific lesion. In order to initially train the network, they use the labelled part of the dataset and, subsequently, they iteratively use a portion of unlabelled ranked patches, starting from the most interesting one, in order to train a network, while asking a medical professional to annotate only those specific patches. The training is stopped when the network has converged, even if not all images were fed to the model, which ultimately ensures that only the most informative ones are actually used during the training.

Khojasteh et al. [140] compared the performance of a CNN, a Discriminative Restricted Boltzmann Machines and the deep features extracted out of a Resnet-50 CNN in combination with several Machine Learning classifiers for detecting exudates in fundus images. The best results were achieved using the deeply learnable features with an SVM classifier. Orlando et al. [102] combined the deep features of a CNN with hand-crafted features into a Random Forest Classifier for detecting early lesions of DR.

7.2.2. Model training with patches

Although segmentation CNNs can be applied on the image at its original resolution, it is very resource-heavy and inefficient. Thus, instead of training the model on the entire image, many researchers crop it in smaller patches instead of just resizing the image at a lower resolution [78, 116, 141–143]. This is necessary when the computational resources are limited or when larger spatial resolution is required, at the expense of a smaller field of view. In Zheng et al. [78], a standalone UNet was trained on fundus patches in order to segment the relevant DR lesions from the images. Two different partitioning approaches were examined for the creation of patches. The first was based on randomly selecting a pixel that was part of an exudate lesion, and taking a patch of 48x48 pixels around it. This approach ensured that the selected patches contained an exudate, but the exudates in several patches may overlap with each other. The second approach was based on iteratively cropping discrete patches out of the image. Although there were no overlapping areas among the various patches, the percentage of them containing an exudate was very small. The model achieved the best F1score of 92.8%, when trained with a mixture of 75% patches from the first approach and 25% from the second one.

7.2.3. Segmentation based on probabilistic output

Several other authors have trained a traditional CNN to generate probabilistic maps, in order to indicate where the lesions are located. In order to avoid redundant boundaries and cluttered pixels around the segmented signs, Khojasteh et al. [76] applied three morphological operations, i.e. closing, opening and finally erosion. However, due to the fact that they generated three output probabilistic maps, one for each of the lesions, it is probable that some pixels will probably belong to more that one lesion. In order to overcome this obstacle, they assigned the pixel with the most probable class, i.e. the one with the highest probability among the three lesions. Both Lam et al. [141] and Benzamin et al. [143] computed the probability of each pixel belonging to each lesion type by scanning each image using a sliding window approach and a traditional CNN. Saha et al. [144] utilized an Encoder-Decoder CNN to segment DR lesions using probabilistic maps. They also included an additional class corresponding to Optic Disc (OD) in order for the network to be able to better differentiate it from exudates.

7.2.4. UNets

As discussed in Section 3, UNets have played a major role in advancing semantic segmentation in many fields. Regarding the fundus imaging and especially Diabetic Retinopathy, several published architectures and AI models have been proposed that are based on a UNet architecture. Yan et al. [145] developed a mutual Global-Local U-Net for segmenting DR related lesions on fundus images. A Global Net and a Local Net are mutually trained using the entire image and its patches respectively, in order for the complete framework to incorporate both local and global information. Sambyal et al. [146] proposed a modified UNet architecture, which utilized a pre-trained ResNet34 as the encoder. For the decoder, 4 UNet blocks with sub-pixel convolutional upsampling with periodic ICNR shuffling [147], which is used to alleviate any checkerboard noise induced by the upsampling procedure.

7.2.5. Segmentation based on attention maps

Gondal et al. [148] used Class Activation Maps in order to visualize possible lesion areas that played a significant role in predicting the severity grade of DR using a classification CNN model. Although attention maps can designate DR lesions, the generated attention maps cannot segment the lesions in detail due to the fact that the main purpose of those networks is to classify the image for its severity. In contrast, such rough estimation of the lesion areas can be interpreted or considered as an estimate of where the network is focusing in order to make its decision. Linking those highlighted areas with true lesions is based on how medical professionals diagnose DR, which is performed by examining the presence of those lesions. However, proving that the model is actually learning such intrinsic characteristics of the data is an open issue, referring to Interpretability and Explainability issues of deep learning models, and cannot be inferred directly. Although attention maps can not be considered accurate and reliable segmentation maps, a very interesting fact was mentioned by Quellec et al. [73], who reported that their classification network managed to optimally detect the various lesions at different training iterations of the model. In particular, the networks (regardless of their architecture) detected the more obvious lesions, i.e. hard and soft exudates, earlier during the training iterations than the more subtle ones, i.e. haemorrhages and then microaneurysms.

7.2.6. DL approaches based on GANs

In an effort to generate synthetic images and enlarge their fundus dataset, many have been utilizing Generative Adversarial Networks (GANs) [78, 103, 109, 149]. A conditional Generative Adversarial Network (cGAN) based on a UNet, was developed in Zheng et al. [78] in order to tackle the problem of the limited and severly imbalanced fundus imaging datasets. By doing so, the authors managed to increase the $F1_{score}$ up to approximately 4% for lesion segmentation. Zhou et al. [109] proposed a GAN model to generate high resolution fundus images that can be manipulated with arbitrary grading and lesion information, which can be used to train both a classification and a segmentation network. GANs have also been used for the segmentation task, apart from the data augmentation. Xiao et al. [68] incorporated a Discriminator network into the pipeline of a traditional segmentation CNN, namely HEDNet, to refine its segmentation, by also minimizing the Discriminator's loss, which was added to the original loss of the CNN.

7.2.7. Per lesion model performance model

Although a similar number of models for segmenting each lesion have been developed, i.e. 19 for exudates [65, 68, 73, 76, 78, 92–94, 121, 130, 134, 138, 140, 143–146, 148, 150] 18 for microaneurysms

[68, 73, 76, 88, 91, 95, 102, 121, 130, 138, 144–146, 148, 150–153], 12 for haemorrhages [68, 73, 76, 94, 102, 121, 138, 144, 145, 148, 150, 151] and 1 for all lesions combined in a 2-class segmentation task [141], their performance is not the same across all lesion types. In particular, the mean accuracy and AUC of the segmentation models regarding the exudate lesions are 97.98 ± 2.35 and 0.684 ± 0.263 respectively. On the other hand, the mean accuracy and AUC regarding the microaneurysms are 92.15 ± 10.17 and 0.565 ± 0.337, while for the haemorrhages these metrics are 93 ± 4.24 and 0.56 ± 0.307 respectively. It becomes evident that exudates are the easiest of the three lesion types to be detected. This is something that can be partly explained by the fact that exudates have a much bigger size than microaneurysms and a distinct yellow color compared to both the haemorrhages and microaneurysms whom color is similar to the retina's vessels' color, as seen in Fig. 2.

7.2.8. Pixel-level imbalancing

Tan et al. [154] noticed that the number of pixels in their dataset that were related to healthy parts of the retina (background) were significantly higher than those related to lesions (29 m background points vs 300k lesion points). Thus, their training procedure was split into 2 phases. During the first phase a smaller but balanced subset of the training set was used (6.4k-9k points each class) to initially train the network, while a much larger one (120k points each class) was used to train the network during phase 2. Eftekhari et al. [152] proposed a 2-stage pipeline for MA detection. In the first stage, the input image is passed through a CNN model to produce a probability map of possible MA regions. Combining that map with the original image, a second CNN detects specific MA and non-MA spots. The authors claim that by doing so, they overcome the serious challenges of the imbalanced datasets, which results in the decrease of the model's false positive rate.

7.2.9. Evaluation performance of published models

Table 8 summarizes the performance results of the segmentation methods as they were reported in the corresponding papers. It should be noted that although we questioned the use of classification-oriented metrics for evaluating the performance of a segmentation model in Section 7.1, we included the performance results of the works using them.

8. Application of DL models in real life clinical settings

Contrary to the findings of a recent review article [20], which clames that *"there are not many methods based on deep learning, and advanced deep learning techniques must be developed in order to solve this problem"*, our analysis indicates that a significant number of DL models has been developed to date and that several of them have actually been employed in the context of clinical decision making in real life environments. What is even more important is that some of them have successfully gone through the regulatory process, having gained approval by relevant international regulatory agencies, such as FDA [108, 155].

Abramoff et al. [108] reported the performance of IDx-DR X2.1 device, which is equipped with a non-mydriatic retinal camera and an automated system for the detection of DR based on CNNs. Although it does not grade the severity of the disease, the system recommends a follow-up with an ophthalmologist regarding the referable DR cases and a 1-year follow-up screening for the non-referable DR cases. In a prospective clinical study, the system was tested on 900 patients on 10 sites [156], and reported a sensitivity of 87.2% and a specificity of 90.7% for more-than-mild DR detection. That led to IDx-DR X2.1 being the first commercial AI device which got the US Food & Drug Administration (FDA) approval in April 2018 [157].

Another FDA-approved commercial AI software for diabetic retinopathy screening is EyeArt [155], which as they claim achieves a sensitivity of 96% and a specificity of 88% for detecting more than mild diabetic retinopathy and a sensitivity of 92% and a specificity of 94% for detecting vision-threatening diabetic retinopathy. EyeArt was also validated on multi-centre dataset of 30405 images from the English Diabetic Eye Screening Programme, achieving high sensitivity for referable DR detection and an acceptable specificity score [125].

Ting et al. [112] proposed a DL-based model to detect referable DR and vision-threatening DR as well as other eye-related diseases. The model was trained and validated on 500.000 images, which were obtained from Singapore National Integrated Diabetic Retinopathy Screening Programme (SiDRP). Their primary objective was to compare the results of their model to several trained and experienced professionals, achieving an AUC of 0.936, sensitivity of 90.5% and specificity of 91.6%, close to that of the professionals who achieved sensitivity of 91.2% and specificity of 99.3% for detecting referable DR. As far as the vision-threatening DR is concerned, the DL-based model achieved an AUC of 0.958, sensitivity of 100% and specificity of 91.1% against the professionals who achieved a sensitivity of 88.5% and a specificity of 99.6%. The model achieved comparable or greater performance in terms of sensitivity but struggled compared to the experts in terms of specificity. In another clinical validation study, Raumviboonsuk et al. [158] reported higher sensitivity but lower specificity scores compared to medical professionals as well. Thus, it is argued that a human-supervised deployment of such a model would be preferable, in order to utilize its high sensitivity but also compensate for its low specificity with the expert's second diagnosis.

Bellemo et al. [159] trained an ensemble of CNNs for diabetic retinopathy grading on fundus images of more than 13.000 patients who had participated in the Singapore Integrated Diabetic Retinopathy Program. Their model was validated through a clinical study on 1574 patients in five urban centres in the Copperbelt province of Zambia. The model achieved clinically acceptable performance in detecting referable DR, vision-threatening DR and diabetic macular edema. These results show that the adoption of a deep learning system for detecting DR is possible and can help developing countries, such those in Africa, where there is lack of expertise and resources, even if the model is trained on a different population. However, their model is trained to detect only the severe non-proliferative and proliferative diabetic retinopathy, leaving out any milder stage of the disease. As they state, a future study which would incorporate all of the stages of diabetic retinopathy is needed, especially because it is equally as important to detect the milder stages of the disease in order to prevent it from deteriorating and ensure early treatment.

9. Discussion

9.1. Data quality and diversity

Many datasets, such as Messidor, IDRiD, etc. consist of high-quality images which were captured under controlled, non-standard conditions (i.e. similar environmental and hardware conditions across captures). Thus, it can be argued that the algorithms trained on such datasets will perform poorly under typical practical situations, where the images may not be directly comparable and the environmental and hardware details may differ. On the other hand, although Kaggle EyePACS and APTOS datasets address these issues and closely resemble a real world scenario, since they consist of images which were captured from a variety of camera models, under various non-typical conditions, the noise which is present due to those variations makes it very difficult for the algorithms to accurately and effectively perform the analysis. However, taking into account those poor-quality images that reflect the actual data, one can develop robust algorithms which can be effective in the clinical practice.

In addition, some datasets, including the large Kaggle EyePACS and APTOS 2019 [46, 47], have been graded by only one expert, which can lead to an annotation bias. Several other datasets either focusing on the DR severity classification [48, 50, 51] or on lesions segmentation [52, 54, 55, 160] have proposed a more complex grading method (majority voting or adjudication on multiple gradings), in order to remove such a

Table 8
Performance of the reviewed segmentation models.

Reference	Dataset	Architecture	Lesion	Accuracy	Sensitivity	Specificity	F1/DICE	IoU	AUC
[76]	DiaretDB1 & e-Ophtha	Custom CNN	Exudate	98%	96%	98%	–	–	–
			Haemorrhage	90%	84%	92%	–	–	–
			Microaneurysm	94%	85%	96%	–	–	–
			Healthy	96%	95%	97%	–	–	–
[130]	IDRiD	CNN based on ResNet	Exudate	99.2%	77.9%	99.6%	–	–	–
			Microaneurysm	99.7%	76.4%	99.8%	–	–	–
	e-Ophtha		Exudate	98.4%	84.6%	98.8%	–	–	–
			Microaneurysm	99.2%	67.2%	99.8%	–	–	–
[141]	Kaggle EyePACS	Inception-V3	All Combined	96%	–	–	–	–	–
[144]	IDRiD	Encoder-Decoder based on VGG16	Soft Exudates	–	–	–	–	–	0.182
			Hard Exudates	–	–	–	–	–	0.550
			Microaneurysms	–	–	–	–	–	0.006
			Haemorrhages	–	–	–	–	–	0.083
[154]	IDRiD	Custom CNN	Healthy	–	95.7%	75.8%	–	–	–
			Exudates	–	87.6%	98.7%	–	–	–
			Microaneurysms	–	62.6%	98.9%	–	–	–
			Haemorrhages	–	46%	98%	–	–	–
[88]	ROC	Custom UNet	Microaneurysms	–	48.5%	–	–	–	–
	DiaretDB1			–	64.1%	–	–	–	–
	e-Ophtha			–	85.9%	–	–	–	–
[143]	IDRiD	Custom CNN	Hard Exudates	96.6%	98.3%	43.4%	–	–	–
[92]	e-Ophtha	Custom CNN	Exudates	91.9%	88.9%	96%	0.926	–	–
[93]	DRiDB	Custom CNN	Exudates	–	78%	–	0.78	–	–
[151]	DiaretDB1	Custom UNet	Red Lesions	–	66.9%	99.8%	0.598	–	–
[148]	DiaretDB1	Custom CNN	Haemorrhages	–	72%	–	–	–	–
			Hard Exudates	–	47%	–	–	–	–
			Soft Exudates	–	71%	–	–	–	–
			Microaneurysms	–	21%	–	–	–	–
[140]	DiaretDB1 e-Ophtha	ResNet	Exudates	98.2%	99%	96%	–	–	–
				97.6%	98%	95%	–	–	–
[73]	DiaretDB1	Custom CNN	Hard Exudates	–	–	–	–	–	0.735
			Soft Exudates	–	–	–	–	–	0.809
			Microaneurysms	–	–	–	–	–	0.5
			Haemorrhages	–	–	–	–	–	0.614
[102]	DiaretDB1	Handcrafted & Deep Features	Red Lesions	–	48.8%	–	–	–	–
	DiaretDB1 & ROC		Small Red Lesions	–	36.8%	–	–	–	–
[78]	e-Ophtha	Custom UNet	Exudates	99.97%	90.94%	99.99%	0.928	–	0.999
	DiaRetDB1			99.97%	93.94%	99.98%	0.925	–	–
	MESSIDOR			99.96%	95.93%	99.99%	0.943	–	–
[94]	IDRiD	UNet	Haemorrhages	–	79.6%	99.9%	0.796	0.67	–
			Hard Exudates	–	84.7%	99.8%	0.815	0.698	–
[152]	ROC	Custom CNN	Microaneurysms	–	76.9%	–	–	–	0.660
	e-Ophtha			–	77.1%	–	–	–	0.637
[153]	DiaretDB1 & Messidor	CNNs Ensemble	Microaneurysms	69.4%	64.6%	88%	–	–	0.834
[95]	ROC	Custom CNN	Microaneurysms	–	39.4%	–	–	–	–
[68]	IDRiD	Custom cGAN	For Every Lesion	–	–	–	0.43–0.46	–	–
[150]	Kaggle EyePACS	Custom CNN	Microaneurysms	–	70.3%	–	–	–	–
			Haemorrhages	–	84.3%	–	–	–	–
			Exudates	–	90.8%	–	–	–	–
[145]	ISBI	Custom UNets	Hard Exudates	–	–	–	–	–	0.889
			Soft Exudates	–	–	–	–	–	0.679
			Microaneurysms	–	–	–	–	–	0.525
			Haemorrhages	–	–	–	–	–	0.703
[138]	IDRiD	Custom CNN	Hard Exudates	–	–	–	–	–	0.795
			Soft Exudates	–	–	–	–	–	0.711
			Microaneurysms	–	–	–	–	–	0.463
			Haemorrhages	–	–	–	–	–	0.637
	DDR		Hard Exudates	–	–	–	–	–	0.555
			Soft Exudates	–	–	–	–	–	0.265
			Microaneurysms	–	–	–	–	–	0.105
			Haemorrhages	–	–	–	–	–	0.359
	e-Ophtha		Hard Exudates	–	–	–	–	–	0.417
			Microaneurysms	–	–	–	–	–	0.169
[121]	IDRiD	Custom UNet	Soft Exudates	–	–	–	–	–	0.9936
			Hard Exudates	–	–	–	–	–	0.9935
			Microaneurysms	–	–	–	–	–	0.9828
			Haemorrhages	–	–	–	–	–	0.9779
[134]	IDRiD	Custom CNN	Hard Exudates	–	78%	–	0.782	–	–
	e-Ophtha			–	51.5%	–	0.496	–	–
[91]	DiaretDB1	Custom CNN	Microaneurysms	87.62%	86.52%	88.73%	0.8742	–	0.9341
	DiaretDB1 (Transfer Learning)			91.4%	91.2%	91.6%	0.913	–	0.962
[65]	DiaretDB1	Custom CNN	Exudates	–	100%	98.4%	–	–	–
[146]	e-Ophtha + IDRiD	Custom UNet	Exudates & MAs	99.9%	99.9%	99.9%	–	–	–

bias and develop a robust and accurate ground truth information for those datasets. Gulshan et al. [110] also proposed that the collected data should be graded multiple times from different professionals, to increase the robustness of the ground truth and in turn the accuracy of the model. Towards that direction, a uniform reference standard should be established to mitigate graders' disagreements [115]. The latter study, showed that an adjudication grading standard was more rigorous, especially in detecting artifacts and missed microaneurysms, than a majority decision protocol.

Poor image quality of the data can affect the training procedure as well as the performance of the model. Subtle signs of retinopathy at an early stage can be easily masked on a low contrast or blurred image. Rakhlin et al. [98] proposed a quality assessment module in their diagnostic pipeline, which discards ungradable images from the dataset. Subsequently, these images are referred to a professional ophthalmologist for examination. Jiang et al. [100] also rejected the low quality images from the final dataset. Li et al. [51] defined the classification as a (6-class) scenary grading problem; the first 5 classes referred to the ICDR grading scale protocol and the 6-th class referred to ungradable images, effectively incorporating the quality assessment module into the deep learning model.

Tan et al. [154] utilized a dataset collected at 11 different clinical sites, using a variety of fundus cameras, and was used to train and test a single CNN model for DR lesion segmentation. During their normalization step, they calculated an A_{score} to assess the quality of the image. Its value was calculated based on the amount of grey pixels, which emerged during normalization of the areas of the image that have little to no illumination. Such dark areas do not provide any meaningful information regarding any anatomical feature of the retina, when they are brightened up, and thus such images were discarded from both the training and the testing set.

On the other hand, Quellec et al. [73] reports that their ensemble model's performance was not largely affected by the image's quality. Nevertheless, increasing the uniformity and consistency among the data, either by controlling the camera's settings and the environmental conditions during the capture or by excluding low-quality images, can improve the model's performance or at least facilitate the training procedure.

Finally, the development of large training and evaluation datasets is one of the many necessary steps towards the development of robust and accurate AI models. However, most of the aforementioned datasets lack sufficient data or suffer from imbalance between their classes. With that in mind, there are several ways to overcome this issue by employing augmentation techniques or generate synthetic data using GANs [109], as well as utilize transfer learning to leverage the knowledge of trained models on large datasets, such as ImageNET [42]. It is also important to increase the diversity of the data regarding their demographics, in order to ensure the model's generazability [53, 86, 110]. Gargeya et al. [53] also proposed to incorporate additional patient metadata, such as genetic factors, duration of diabetes, hemoglobin A1C value, and other clinical data that may influence their risk for developing retinopathy. It also may be of interest to include specific information related to explicit lesion features within the classification models [110]. Doing so, the AI model may yield insightful correlations into underlying DR risk factors and potentially increasing the diagnostic performance.

9.2. AI acceptance and clinical integration

Artificial Intelligence (AI)- and especially Deep Learning (DL)-based methods hold promise for improving and accelerating healthcare. However, there are several key constraints that need to be addressed in order to facilitate AI's adoption in clinical settings [161]. Apart from the traditional methods that are used to assess the model's performance, i.e. accuracy metrics, several others are proposed as important elements towards the acceptance of AI models through regulatory processes [162]. The progress from traditional machine learning approaches to deep learning ones, although it has improved the performance of such analyses, has also been accompanied by a lack of explainability and transparency. The interpretability of such models is however a crucial element affecting their acceptance and integration in the clinical practice. The clinical operator needs to understand the model's decision process which should ideally provide explanations regarding its predictions (e.g. why these predictions were made and what alternatives were considered). Regarding DR, several researchers have generated evidence heatmaps, in an attempt to aggregate the importance of each pixel to the prediction across the several network's layers [73, 100, 117]. Such visualizations allow the clinicians to determine whether the model bases its prediction on relevant clinical features, which in the case of DR would include exudates, microaneurysms and haemorrhages as previously discussed. In addition, two other crucial elements regarding the robustness and reliability of the models, need to be properly addressed prior to clinical integration. These terms ultimately refer to the need of the models to consistently perform accurately across expected variations encountered in the clinical environment, including variations regarding data collected from multiple centres or machines from various vendors.

10. Conclusion

Diabetic retinopathy is a serious complication of diabetes mellitus, leading to progressive damage and even blindness of the retina. Its early detection and treatment is important in order to prevent its deterioration and the retina's damage. The interest in applying deep learning in detecting diabetic retinopathy has increased during the past years and as several DL systems evolve and become integrated into the clinical practice, they will enable the clinicians to treat the patients in need more effectively and efficiently. This article presents the current state of research regarding the application of deep learning in diagnosing diabetic retinopathy. Although deep learning has paved the way for more accurate diagnosis and treatment, further improvements are still necessary regarding performance, interpretability and trustworthiness from ophthalmologists.

Author contributions

All authors have read and approved the manuscript, and each author has participated sufficiently in developing the manuscript. N.T., D.T, and G.M. contributed to the literature review and analysis of the study and drafting the manuscript. E.K., and O.B. contributed to the clinical aspects and drafting the manuscript. A.B., F.S., A.S., D.I.F., and K.M reviewed the manuscript and contributed in the interpretation of the findings.

Funding

This work was partially supported by the H2020 specific targeted research project SeeFar: Smart glasses for multifacEted visual loss mitigation and chronic disEase prevention indicator for healthier, saFer, and more productive workplAce foR ageing population. (H2020-SC1-DTH-2018-1, GA No 826429) (www.see-far.eu). This paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Declaration of competing interest

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104599.

References

[1] International Diabetes Federation. International diabetes federation diabetes atlas, ninth ed.https://www.diabetesatlas.org/en/.

[2] Alicia J. Jenkins, Mugdha V. Joglekar, Anandwardhan A. Hardikar, Anthony C. Keech, David N. O'Neal, S. Andrzej, Januszewski, Biomarkers in diabetic retinopathy, Rev. Diabet. Stud.: Reg. Dev. Stud. 12 (1–2) (2015) 159.

[3] Mohsen Janghorbani, Raymond B. Jones, Simon P. Allison, Incidence of and risk factors for proliferative retinopathy and its association with blindness among diabetes clinic attenders, Ophthalmic Epidemiol. 7 (4) (2000) 225–241.

[4] J.W. Yau, S.L. Rogers, R. Kawasaki, E.L. Lamoureux, J.W. Kowalski, T. Bek, S. J. Chen, J.M. Dekker, A. Fletcher, J. Grauslund, Meta-analysis for eye disease [meta-eye] study group. *Global prevalence and major risk factors of diabetic retinopathy*, Diabetes Care 35 (3) (2012) 556–564.

[5] Jonathan E. Shaw, Richard A. Sicree, Paul Z. Zimmet, Global estimates of the prevalence of diabetes for 2010 and 2030, Diabetes Res. Clin. Pract. 87 (1) (2010) 4–14.

[6] Wenying Yang, Juming Lu, Jianping Weng, Weiping Jia, Linong Ji, Jianzhong Xiao, Zhongyan Shan, Jie Liu, Haoming Tian, Qiuhe Ji, Prevalence of diabetes among men and women in China, N. Engl. J. Med. 362 (12) (2010) 1090–1101.

[7] Safi Hamid, Sare Safi, Ali Hafezi-Moghadam, Ahmadieh Hamid, Early detection of diabetic retinopathy, Surv. Ophthalmol. 63 (5) (2018) 601–608.

[8] Scottish Intercollegiate Guideline Network, Management of Diabetes: a National Clinical Guideline, Scottish Intercollegiate Guidelines Network, Edinburgh, 2014.

[9] H Bresnick George, Dana B. Mukamel, John C. Dickinson, David R. Cole, A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy, Ophthalmology 107 (1) (2000) 19–24.

[10] Linda Hill, E. Lydia, Makaroff. Early detection and timely treatment can prevent or delay diabetic retinopathy, Diabetes Res. Clin. Pract. 120 (October 2016) 241–243, https://doi.org/10.1016/j.diabres.2016.09.004.

[11] Thanh Tan Nguyen, Jie Jin Wang, A Richey Sharrett, FM Amirul Islam, Ronald Klein, Barbara EK. Klein, Mary Frances Cotch, Tien Yin Wong, Relationship of retinal vascular caliber with diabetes and retinopathy: the multi-ethnic study of atherosclerosis (mesa), Diabetes Care 31 (3) (2008) 544–549.

[12] Judith Lechner, Olivia E O'Leary, Alan W Stitt, The pathology associated with diabetic retinopathy, Vis. Res. 139 (7–14) (2017).

[13] Changyow C. Kwan, Amani A. Fawzi, Imaging and biomarkers in diabetic macular edema and diabetic retinopathy, Curr. Diabetes Rep. 19 (10) (2019) 1–10.

[14] Rajiv Raman, Suganeswari Ganesan, Swakshyar Saumya Pal, Vaitheeswaran Kulothungan, Tarun Sharma, Prevalence and risk factors for diabetic retinopathy in rural India. sankara nethralaya diabetic retinopathy epidemiology and molecular genetic study iii (sn-dreams iii), report no 2, BMJ Open Diabet. Res Care 2 (1) (2014).

[15] Serge Resnikoff, William Felch, Tina-Marie Gauthier, Spivey Bruce, The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners, Br. J. Ophthalmol. 96 (6) (2012) 783–787.

[16] P.I. Burgess, I.J.C. MacCormick, S.P. Harding, A. Bastawrous, Nav Beare, Paul Garner, Epidemiology of diabetic retinopathy and maculopathy in africa: a systematic review, Diabet. Med. 30 (4) (2013) 399–412.

[17] Sobha Sivaprasad, Bhaskar Gupta, Roxanne Crosby-Nwaobi, Jennifer Evans, Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective, Surv. Ophthalmol. 57 (4) (2012) 347–370.

[18] Andrzej Grzybowski, Piotr Brona, Gilbert Lim, Paisan Ruamviboonsuk, Gavin SW. Tan, Michael Abramoff, SW Ting Daniel, Artificial intelligence for diabetic retinopathy screening: a review, Eye 34 (3) (2020) 451–460.

[19] Maryam Badar, Muhammad Haris, Anam Fatima, Application of deep learning for retinal image analysis: a review, Comput. Sci. Rev. 35 (2020) 100203.

[20] Norah Asiri, Muhammad Hussain, Fadwa Al Adel, Nazih Alzaidi, Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey, Artif. Intell. Med. 99 (2019) 101701.

[21] Md Mohaimenul Islam, Hsuan-Chia Yang, Tahmina Nasrin Poly, Wen-Shan Jian, Yu-Chuan Jack Li, Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis, Comput. Methods Progr. Biomed. 191 (2020) 105320.

[22] Katrine B. Nielsen, Mie L. Lautrup, Jakob KH. Andersen, R Savarimuthu Thiusius, Jakob Grauslund, Deep learning–based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance, Ophthalmol. Retina 3 (4) (2019) 294–304.

[23] Noemi Lois, Rachel V. McCarter, Christina O'Neill, J. Reinhold, Medina, Alan W Stitt, Endothelial progenitor cells in diabetic retinopathy, Front. Endocrinol. 5 (44) (2014).

[24] Alan W Stitt, Christina L. O'Neill, Michelle T. O'Doherty, Desmond B. Archer, Tom A. Gardiner, J. Reinhold, Medina, Vascular stem cells and ischaemic retinopathies, Prog. Retin. Eye Res. 30 (3) (2011) 149–166.

[25] Ronald Klein, B.E. Klein, Scot E. Moss, The Wisconsin epidemiological study of diabetic retinopathy: a review, Diabetes Metab. Rev. 5 (7) (1989) 559–570.

[26] L.Z. Heng, O. Comyn, T. Peto, C. Tadros, E. Ng, S. Sivaprasad, P.G. Hykin, Diabetic retinopathy: pathogenesis, clinical grading, management and future developments, Diabet. Med. 30 (6) (2013) 640–650.

[27] Early Treatment Diabetic Retinopathy Study Research Group, Early treatment diabetic retinopathy study design and baseline patient characteristics: etdrs report number 7, Ophthalmology 98 (5) (1991) 741–756.

[28] C.P. Wilkinson, Frederick L. Ferris III, Ronald E. Klein, Paul P. Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R. Pararajasegaram, Juan T. Verdaguer, Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, Ophthalmology 110 (9) (2003) 1677–1682.

[29] T.J. Verdaguer, Screening para retinopatia diabetica en latino America. resultados, Rev Soc Brasil Retina Vitreo 4 (14–5) (2001).

[30] Masatoshi Fukuda, Clinical arrangement of classification of diabetic retinopathy, Tohoku J. Exp. Med. 141 (Suppl) (1983) 331–335.

[31] Diabetic Retinopathy Working Party, Management of Diabetic Retinopathy: Clinical Practice Guidelines (June 1997), NHMRC, Canberra, 1997, pp. 1–94.

[32] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[33] Laurent Itti, Christof Koch, Ernst Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[34] Ronald A. Rensink, The dynamic representation of scenes, Vis. Cognit. 7 (1–3) (2000) 17–42.

[35] Maurizio Corbetta, Gordon L. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nat. Rev. Neurosci. 3 (3) (2002) 201–215.

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, Xiaoou Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[38] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[39] Jongchan Park, Sanghyun Woo, Joon-Young Lee, In So Kweon, Bam: bottleneck attention module, 2018 arXiv preprint arXiv:1807.06514.

[40] Jean Pouget-Abadie Ian J Goodfellow, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial networks, 2014 *arXiv preprint arXiv:1406.2661,*.

[41] Ian Goodfellow, Nips 2016 Tutorial: Generative Adversarial Networks, 2016 arXiv preprint arXiv:1701.00160.

[42] Deng Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[43] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Yoshua Bengio, Deep Learning, 1, MIT press Cambridge, 2016.

[44] Tongxue Zhou, Ruan Su, Stéphane Canu, A review: deep learning for medical image segmentation using multi-modality fusion, Array 3 (2019) 100004.

[45] Samuel Budd, Emma C. Robinson, Bernhard Kainz, A Survey on Active Learning and Human-In-The-Loop Deep Learning for Medical Image Analysis, Medical Image Analysis, 2021, p. 102062.

[46] EyePACS, Diabetic retinopathy detection. www.kaggle.com/c/diabetic-retinopathy-detection. (Accessed 25 February 2021).

[47] Aravind Eye Hospital, APTOS 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed on 25/02/2021.

[48] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, Jean-Claude Klein, Feedback on a publicly distributed database: the messidor database, Image Anal. Stereol. 33 (3) (August 2014) 231–234, https://doi.org/10.5566/ias.1155. ISSN 1854–5165.

[49] Michael D. Abràmoff, James C. Folk, Dennis P. Han, Jonathan D. Walker, David F. Williams, Stephen R. Russell, Massin Pascale, Beatrice Cochener, Philippe Gain, Li Tang, Mathieu Lamard, Daniela C. Moga, Gwénolé Quellec, Meindert Niemeijer, Automated analysis of retinal images for detection of referable diabetic retinopathy, JAMA Ophthalmol. 131 (3) (mar 2013) 351, https://doi.org/10.1001/jamaophthalmol.2013.1743. ISSN 2168–6165.

[50] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau, Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research, Data 3 (3) (2018), https://doi.org/10.3390/data3030025. ISSN 23065729.

[51] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, Hong Kang, Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, Inf. Sci. 501 (oct 2019) 511–522, https://doi.org/10.1016/J.INS.2019.06.011. ISSN 0020–0255.

[52] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, A. Chabouis, TeleOphta: Machine learning and image processing methods for teleophthalmology, IRBM 34 (2) (apr 2013) 196–203, https://doi.org/10.1016/J.IRBM.2013.01.010. ISSN 1959–0318.

[53] Rishab Gargeya, Theodore Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (jul 2017) 962–969, https://doi.org/10.1016/j.ophtha.2017.02.008. ISSN 15494713.

[54] R.V.J.P.H. Kälviäinen, H. Uusitalo, Diaretdb1 diabetic retinopathy database and evaluation protocol, in: Medical Image Understanding and Analysis, 2007, Citeseer, 2007, p. 61.

[55] Pavle Prentasic, Sven Loncaric, Zoran Vatavuk, Goran Bencic, Marko Subasic, Tomislav Petkovic, Lana Dujmovic, Maja Malenica-Ravlic, Nikolina Budimlija,

Raseljka Tadic, Diabetic retinopathy image database(DRiDB): a new database for diabetic retinopathy screening programs research, in: 2013 8th *International Symposium On Image And Signal Processing And Analysis (ISPA)*, IEEE, sep 2013, ISBN 978-953-184-194-8, pp. 711–716, https://doi.org/10.1109/ISPA.2013.6703830.

[56] Meindert Niemeijer, Bram Van Ginneken, Michael J. Cree, Atsushi Mizutani, Gwénolé Quellec, Clara I Sánchez, Bob Zhang, Roberto Hornero, Mathieu Lamard, Chisako Muramatsu, Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs, IEEE Trans. Med. Imag. 29 (1) (2009) 185–195.

[57] A.D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imag. 19 (3) (March 2000) 203–210, https://doi.org/10.1109/42.845178. ISSN 0278-0062.

[58] Joes Staal, Michael D. Abràmoff, Meindert Niemeijer, Max A. Viergever, Bram Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imag. 23 (4) (2004) 501–509.

[59] Zhuo Zhang, Feng Shou Yin, Liu Jiang, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, Tien Yin Wong, ORIGA^{-light}: an online retinal fundus image database for glaucoma analysis and research, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE, aug 2010, ISBN 978-1-4244-4123-5, pp. 3065–3068, https://doi.org/10.1109/IEMBS.2010.5626137.

[60] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G. Owen, Sarah A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 59 (9) (2012) 2538–2548.

[61] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, Georg Michelson, Robust vessel segmentation in fundus images, Int. J. Biomed. Imag. (2013) 2013.

[62] Wei Zhang, Jie Zhong, Shijun Yang, Zhentao Gao, Junjie Hu, Yuanyuan Chen, Yi Zhang, Automated identification and grading system of diabetic retinopathy using deep neural networks, Knowl. Base Syst. 175 (jul 2019) 12–25, https://doi.org/10.1016/j.knosys.2019.03.016. ISSN 9507051.

[63] Z. Zhao, K. Zhang, X. Hao, J. Tian, M.C. Heng Chua, L. Chen, X. Xu, Bira-net: bilinear attention net for diabetic retinopathy grading, in: 2019 IEEE International Conference on Image Processing, ICIP, 2019, pp. 1385–1389, https://doi.org/10.1109/ICIP.2019.8803074.

[64] Darshit Doshi, Aniket Shenoy, Deep Sidhpura, Prachi Gharpure, Diabetic retinopathy detection using deep convolutional neural networks, in: 2016 *International Conference On Computing, Analytics And Security Trends (CAST)*, IEEE, dec 2016, ISBN 978-1-5090-1338-8, pp. 261–266, https://doi.org/10.1109/CAST.2016.7914977.

[65] Kemal Adem, Exudate detection for diabetic retinopathy with circular hough transformation and convolutional neural networks, Expert Syst. Appl. 114 (2018) 289–295.

[66] Carson Lam, Darvin Yi, Margaret Guo, Tony Lindsey, Automated detection of diabetic retinopathy using deep learning, AMIA Joint Summits Transl. Sci. Proc. (2018) 147–155. ISSN 2153–4063.

[67] Lizong Zhang, Shuxin Feng, Guiduo Duan, Ying Li, Guisong Liu, Detection of microaneurysms in fundus images based on an attention mechanism, Genes 10 (10) (2019) 817.

[68] Qiqi Xiao, Jiaxu Zou, Muqiao Yang, Alex Gaudio, Kris Kitani, Asim Smailagic, Pedro Costa, Min Xu, Improving lesion segmentation for diabetic retinopathy using adversarial learning, in: International Conference on Image Analysis and Recognition, Springer, 2019, pp. 333–344.

[69] Bo Wu, Weifang Zhu, Fei Shi, Shuxia Zhu, Xinjian Chen, Automatic detection of microaneurysms in retinal fundus images, Comput. Med. Imag. Graph. 55 (2017) 106–112.

[70] Misgina Tsighe Hagos, Shri Kant, Transfer Learning Based Detection of Diabetic Retinopathy from Small Dataset, 2019 arXiv preprint arXiv:1905.07203.

[71] Sheikh Muhammad Saiful Islam, Md Mahedi Hasan, Sohaib Abdullah, Deep Learning Based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images, 2018 arXiv preprint arXiv:1812.10595.

[72] HaiQuan Chen, XiangLong Zeng, Yuan Luo, WenBin Ye, Detection of diabetic retinopathy using deep neural network, in: 2018 IEEE 23rd International Conference On Digital Signal Processing (DSP, IEEE, nov 2018, ISBN 978-1-5386-6811-5, pp. 1–5, https://doi.org/10.1109/ICDSP.2018.8631882.

[73] Gwenolé Quellec, Katia Charrière, Yassine Boudi, Béatrice Cochener, Mathieu Lamard, Deep image mining for diabetic retinopathy screening, Med. Image Anal. 39 (2017) 178–193, https://doi.org/10.1016/j.media.2017.04.012. ISSN 13618423.

[74] María A. Bravo, Pablo A. Arbeláez, Automatic diabetic retinopathy classification, in: 13th International Conference on Medical Information Processing and Analysis, 10572, International Society for Optics and Photonics, 2017, p. 105721E.

[75] Gabriel García, Jhair Gallardo, Antoni Mauricio, Jorge López, Christian Del Carpio, Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 635–642, https://doi.org/10.1007/978-3-319-68612-7\text{_}72.

[76] Parham Khojasteh, Behzad Aliahmad, Dinesh K. Kumar, Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms, BMC Ophthalmol. 18 (1) (2018) 1–13.

[77] Sohini Roychowdhury, Dara D. Koozekanani, K Parhi Keshab, Dream: diabetic retinopathy analysis using machine learning, IEEE J. Biomed. Health Inform. 18 (5) (2013) 1717–1728.

[78] Rui Zheng, Lei Liu, Shulin Zhang, Chun Zheng, Filiz Bunyak, Ronald Xu, Bin Li, Mingzhai Sun, Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network, Biomed. Opt Express 9 (10) (2018) 4863–4878.

[79] Ratul Ghosh, Kuntal Ghosh, Sanjit Maitra, Automatic detection and classification of diabetic retinopathy stages using CNN, in: 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2017, ISBN 978-1-5090-2797-2, pp. 550–554, https://doi.org/10.1109/SPIN.2017.8050011.

[80] Sehrish Qummar, Fiaz Gul Khan, Sajid Shah, Ahmad Khan, Shahaboddin Shamshirband, Zia Ur Rehman, Iftikhar Ahmed Khan, Waqas Jadoon, A deep learning ensemble approach for diabetic retinopathy detection, IEEE Acc. (2019), https://doi.org/10.1109/access.2019.2947484, 1–1.

[81] Gen-Min Lin, Mei-Juan Chen, Chia-Hung Yeh, Yu-Yang Lin, Heng-Yu Kuo, Min-Hui Lin, Ming-Chin Chen, Shinfeng D. Lin, Ying Gao, Anran Ran, Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy, J. Ophthalmol. 2018 (2018).

[82] Harry Pratt, Frans Coenen, Deborah M. Broadbent, Simon P. Harding, Yalin Zheng, Convolutional neural networks for diabetic retinopathy, in: Procedia Computer Science, vol. 90, Elsevier, jan 2016, pp. 200–205, https://doi.org/10.1016/j.procs.2016.07.014.

[83] Shaohua Wan, Yan Liang, Yin Zhang, Deep convolutional neural networks for diabetic retinopathy detection by image classification, Comput. Electr. Eng. 72 (nov 2018) 274–282, https://doi.org/10.1016/J.COMPELECENG.2018.07.042. ISSN 0045–7906.

[84] Manoj Raju, Pagidimarri Venkatesh, Barreto Ryan, Amrit Kadam, Vamsichandra Kasivajjala, Arun Aswath, Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy, in: Studies in Health Technology and Informatics, 245, 2017, ISBN 9781614998297, pp. 559–563, https://doi.org/10.3233/978-1-61499-830-3-559.

[85] Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, Pheng-Ann Heng, Canet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading, IEEE Trans. Med. Imag. 39 (5) (2020) 1483–1493.

[86] Yogesan Kanagasingam, Di Xiao, Janardhan Vignarajan, Amita Preetham, Mei-Ling Tay-Kearney, Ateev Mehrotra, Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care, JAMA Network Open 1 (5) (sep 2018), e182665, https://doi.org/10.1001/jamanetworkopen.2018.2665. ISSN 2574–3805.

[87] Piotr Chudzik, Somshubra Majumdar, Francesco Caliva, Bashir Al-Diri, Andrew Hunter, Exudate segmentation using fully convolutional neural networks and inception modules, in: Medical Imaging 2018: Image Processing, 10574, International Society for Optics and Photonics, 2018, p. 1057430.

[88] Piotr Chudzik, Somshubra Majumdar, Francesco Calivá, Bashir Al-Diri, Andrew Hunter, Microaneurysm detection using fully convolutional neural networks, Comput. Methods Progr. Biomed. 158 (2018) 185–192.

[89] Elaheh Imani, Hamid-Reza Pourreza, A novel method for retinal exudate segmentation using signal separation algorithm, Comput. Methods Progr. Biomed. 133 (2016) 195–205.

[90] Rmbtp Udhaya Sankar, R. Vijai, R.M. Balajee, Detection and classification of diabetic retinopathy in fundus images using neural network, Int. Res. J. Eng. Technol 5 (4) (2018) 2630–2635.

[91] Juan Shan, Lin Li, A deep learning method for microaneurysm detection in fundus images, in: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), IEEE, 2016, pp. 357–358.

[92] Shuang Yu, Di Xiao, Yogesan Kanagasingam, Exudate detection for diabetic retinopathy with convolutional neural networks, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 1744–1747.

[93] Pavle Prentašić, Sven Lončarić, Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion, Comput. Methods Progr. Biomed. 137 (2016) 281–292.

[94] Ashutosh Kushwaha, P. Balamurugan, Classifying diabetic retinopathy images using induced deep region of interest extraction, in: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2019, pp. 1–6.

[95] Umit Budak, Abdulkadir Şengür, Yanhui Guo, Yaman Akbulut, A novel microaneurysms detection approach based on convolutional neural networks with reinforcement sample learning algorithm, Health Inf. Sci. Syst. 5 (1) (2017) 14.

[96] Shu-I Pao, Hong-Zin Lin, Ke-Hung Chien, Ming-Cheng Tai, Jiann-Torng Chen, Gen-Min Lin, Detection of diabetic retinopathy using bichannel convolutional neural network, J. Ophthalmol. (2020) 2020.

[97] Igi Ardiyanto, Hanung Adi Nugroho, Ratna Lestari Budiani Buana, Deep learning-based Diabetic Retinopathy assessment on embedded system, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, jul 2017, ISBN 978-1-5090-2809-2, pp. 1760–1763, https://doi.org/10.1109/EMBC.2017.8037184.

[98] Rakhlin Alexander, Diabetic Retinopathy Detection through Integration of Deep Learning Classification Framework, bioRxiv, jun 2018, p. 225508, https://doi.org/10.1101/225508.

[99] Xiaoliang Wang, Yongjin Lu, Yujuan Wang, Wei-Bang Chen, Diabetic retinopathy stage classification using convolutional neural networks, in: 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, jul

2018, ISBN 978-1-5386-2659-7, pp. 465–471, https://doi.org/10.1109/IRI.2018.00074.
[100] Hongyang Jiang, Yang Kang, Mengdi Gao, Dongdong Zhang, Ma He, Wei Qian, An interpretable ensemble deep learning model for diabetic retinopathy disease classification, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, jul 2019, ISBN 978-1-5386-1311-5, pp. 2045–2048, https://doi.org/10.1109/EMBC.2019.8857160.
[101] Jordi de La Torre, Aida Valls, Domenec Puig, A deep learning interpretable classifier for diabetic retinopathy disease grading, Neurocomputing 396 (2020) 465–476.
[102] José Ignacio Orlando, Prokofyeva Elena, Mariana del Fresno, Matthew B. Blaschko, An ensemble deep learning based approach for red lesion detection in fundus images, Comput. Methods Progr. Biomed. 153 (2018) 115–127.
[103] K Pujitha Appan, Jayanthi Sivaswamy, Retinal image synthesis for cad development, in: International Conference Image Analysis and Recognition, Springer, 2018, pp. 613–621.
[104] Nobuyuki Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybernet. 9 (1) (1979) 62–66.
[105] J.L. Starck, Michael Elad, David L. Donoho, Image decomposition via the combination of sparse representations and a variational approach, IEEE Trans. Image Process. 14 (10) (2005) 1570–1582.
[106] Gitta Kutyniok, Jakob Lemvig, Wang Q. Lim, Compactly supported shearlets, in: Approximation Theory XIII: San Antonio, Springer, 2010, pp. 163–186, 2012.
[107] Minh N. Do, Martin Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Trans. Image Process. 14 (12) (2005) 2091–2106.
[108] Michael David Abràmoff, Yiyue Lou, Erginay Ali, Clarida Warren, Amelon Ryan, James C. Folk, Meindert Niemeijer, Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, Invest. Opthalmol. Vis. Sci. 57 (13) (oct 2016) 5200, https://doi.org/10.1167/iovs.16-19964. ISSN 1552–5783.
[109] Yi Zhou, Boyang Wang, Xiaodong He, Shanshan Cui, Ling Shao, Dr-gan: conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images, IEEE J. Biomed. Health Inform. (2020).
[110] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, Dale R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, J. Am. Med. Assoc. 316 (22) (dec 2016) 2402, https://doi.org/10.1001/jama.2016.17216. ISSN 0098–7484.
[111] Kele Xu, Dawei Feng, Haibo Mi, Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image, Molecules 22 (12) (nov 2017) 2054, https://doi.org/10.3390/molecules22122054. ISSN 1420–3049.
[112] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, Tien Yin Wong, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, L. Ecosse, Y. Ian, Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, Tien Yin Wong, Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, J. Am. Med. Assoc. 318 (22) (dec 2017) 2211.
[113] K. Zhou, Z. Gu, W. Liu, W. Luo, J. Cheng, S. Gao, J. Liu, Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, July 2018, pp. 2724–2727, https://doi.org/10.1109/EMBC.2018.8512828.
[114] Fangjun Li, Dongfeng Yuan, Mingqiang Zhang, Cong Liang, Xiaotian Zhou, Haixia Zhang, Multi-scale stepwise training strategy of convolutional neural networks for diabetic retinopathy severity assessment, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–5.
[115] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, Dale R. Webster, Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy, Ophthalmology 125 (8) (aug 2018) 1264–1272, https://doi.org/10.1016/j.ophtha.2018.01.034. ISSN 1616420.
[116] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, Xiaogang Wang, Zoom-in-net: deep mining lesions for diabetic retinopathy detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 267–275.
[117] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Barb Scott, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, Dale R. Webster, Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (4) (apr 2019) 552–564, https://doi.org/10.1016/J.OPHTHA.2018.11.016. ISSN 0161–6420.
[118] Tom Fawcett, An introduction to roc analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.
[119] Laith Alzubaidi, Jinglan Zhang, J Humaidi Amjad, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, Laith Farhan, Review of deep learning: concepts, cnn architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021) 1–74.
[120] Zhiwen Lin, Ruoqian Guo, Yanjie Wang, Bian Wu, Tingting Chen, Wenzhe Wang, Danny Z. Chen, Jian Wu, A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 74–82.
[121] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, Ling Shao, Collaborative learning of semi-supervised segmentation and classification for medical images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2079–2088.
[122] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, in: 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR, 2014.
[123] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[124] Michelle Yuen Ting Yip, Wei Lim Zhan, Gilbert Lim, Nguyen Duc Quang, Haslina Hamzah, Jinyi Ho, Valentina Bellemo, Yuchen Xie, Xin Qi Lee, Mong Li Lee, Enhanced detection of referable diabetic retinopathy via dcnns and transfer learning, in: Asian Conference on Computer Vision, Springer, 2018, pp. 282–288.
[125] Heydon Peter, Catherine Egan, Louis Bolter, Ryan Chambers, John Anderson, Aldington Steve, M Stratton Irene, Peter Henry Scanlon, Laura Webster, Samantha Mann, Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients, Br. J. Ophthalmol. 105 (5) (2021) 723–728.
[126] Daniel P. Huttenlocher, Gregory A. Klanderman, William J. Rucklidge, Comparing images using the hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 15 (9) (1993) 850–863.
[127] Abdel Aziz Taha, Hanbury Allan, Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool, BMC Med. Imag. 15 (1) (2015) 1–28.
[128] Orhun Utku Aydin, Abdel Aziz Taha, Hilbert Adam, Ahmed A. Khalil, Ivana Galinovic, Jochen B. Fiebach, Dietmar Frey, Vince Istvan Madai, On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking, Euro. Radiol. Exp. 5 (1) (2021) 1–7.
[129] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N. Conrad, Esha Datta, Dávid Gergely, Benjamin De Leener, et al., Spinal cord grey matter segmentation challenge, Neuroimage 152 (2017) 312–329.
[130] Jie Xue, Shuo Yan, Jianhua Qu, Feng Qi, Chenggong Qiu, Hongyan Zhang, Meirong Chen, Tingting Liu, Dengwang Li, Xiyu Liu, Deep membrane systems for multitask segmentation in diabetic retinopathy, Knowl. Base Syst. 183 (2019) 104887.
[131] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[132] Xingyi Zhang, Yanjun Liu, Bin Luo, Linqiang Pan, Computational power of tissue p systems for generating control languages, Inf. Sci. 278 (2014) 285–297.
[133] Bosheng Song, Linqiang Pan, Mario J. Pérez-Jiménez, Cell-like p systems with channel states and symport/antiport rules, IEEE Trans. NanoBioscience 15 (6) (2016) 555–566.
[134] Song Guo, Tao Li, Kai Wang, Chan Zhang, Hong Kang, A lightweight neural network for hard exudate segmentation of fundus image, in: International Conference on Artificial Neural Networks,, Springer, 2019, pp. 189–199.
[135] Liang-Chieh Chen, Yukun Zhu, Papandreou George, Florian Schroff, Hartwig Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[136] Juan Mo, Lei Zhang, Yangqin Feng, Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks, Neurocomputing 290 (2018) 161–171.
[137] Saining Xie, Zhuowen Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[138] Song Guo, Tao Li, Hong Kang, Li Ning, Yujun Zhang, Kai Wang, L-seg: an end-to-end unified framework for multi-lesion segmentation of fundus images, Neurocomputing 349 (2019) 52–63.
[139] Sebastian Otálora, Oscar Perdomo, Fabio González, Henning Müller, Training deep convolutional neural networks with active learning for exudate classification in eye fundus images, in: Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, 2017, pp. 146–154.
[140] Parham Khojasteh, Leandro Aparecido Passos Júnior, Tiago Carvalho, Edmar Rezende, Behzad Aliahmad, João Paulo Papa, Dinesh Kant Kumar, Exudate detection in fundus images using deeply-learnable features, Comput. Biol. Med. 104 (2019) 62–69.
[141] Carson Lam, Caroline Yu, Laura Huang, Daniel Rubin, Retinal lesion detection with deep learning using image patches, Invest. Ophthalmol. Vis. Sci. 59 (1) (2018) 590–596.
[142] Cao Wen, Juan Shan, Nicholas Czarnek, Li Lin, Microaneurysm detection in fundus images using small image patches and machine learning methods, in: 2017

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2017, pp. 325–331.
[143] Avula Benzamin, Chandan Chakraborty, Detection of hard exudates in retinal fundus images using deep learning, in: 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, 2018, pp. 465–469.
[144] Oindrila Saha, Rachana Sathish, Debdoot Sheet, Fully convolutional neural network for semantic segmentation of anatomical structure and pathologies in colour fundus images associated with diabetic retinopathy, 2019 arXiv preprint arXiv:1902.03122.
[145] Zizheng Yan, Xiaoguang Han, Changmiao Wang, Yuda Qiu, Zixiang Xiong, Shuguang Cui, Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 597–600.
[146] Nitigya Sambyal, Poonam Saini, Rupali Syal, Varun Gupta, Modified u-net architecture for semantic segmentation of diabetic retinopathy images, Biocyber. Biomed. Eng. 40 (3) (2020) 1094–1109.
[147] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, Wenzhe Shi, Checkerboard Artifact Free Sub-pixel Convolution: A Note on Sub-pixel Convolution, Resize Convolution and Convolution Resize, 2017 arXiv preprint arXiv:1707.02937.
[148] Waleed M Gondal, Jan M. Köhler, René Grzeszick, Gernot A. Fink, Michael Hirsch, Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 2069–2073.
[149] Paolo Andreini, Simone Bonechi, Monica Bianchini, Alessandro Mecocci, Scarselli Franco, Andrea Sodi, A Two Stage gan for High Resolution Retinal Image Generation and Segmentation, 2019 arXiv preprint arXiv:1907.12296.
[150] Yehui Yang, Tao Li, Wensi Li, Haishan Wu, Wei Fan, Wensheng Zhang, Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 533–540.
[151] Clément Playout, Renaud Duval, Farida Cheriet, A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 101–108.
[152] Noushin Eftekhari, Hamid-Reza Pourreza, Mojtaba Masoudi, Kamaledin Ghiasi-Shirazi, Ehsan Saeedi, Microaneurysm detection in fundus images using a two-step convolutional neural network, Biomed. Eng. Online 18 (1) (2019) 67.
[153] Balazs Harangi, Janos Toth, Andras Hajdu, Fusion of deep convolutional neural networks for microaneurysm detection in color fundus images, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 3705–3708.
[154] Jen Hong Tan, Hamido Fujita, Sobha Sivaprasad, V Bhandary Sulatha, A. Krishna Rao, Chua Chua Kuang, U Rajendra Acharya, Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network, Inf. Sci. 420 (2017) 66–76.
[155] Eyenuk, Eyeart ai eye screening system. https://www.eyenuk.com/. Accessed on 25/02/21.
[156] Michael D. Abràmoff, Philip T. Lavin, Michele Birch, Nilay Shah, C Folk James, Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices, NPJ Digital Med. 1 (1) (2018) 1–8.
[157] FDA, Fda's approval of idx-dr. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye. Accessed on 25/02/2021.
[158] Paisan Raumviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, JL Campana Bilson, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, Deep Learning vs. Human Graders for Classifying Severity Levels of Diabetic Retinopathy in a Real-World Nationwide Screening Program, 2018 arXiv preprint arXiv:1810.08290.
[159] Valentina Bellemo, W Lim Zhan, Gilbert Lim, Quang D. Nguyen, Yuchen Xie, Michelle YT. Yip, Haslina Hamzah, Jinyi Ho, Q Lee Xin, Wynne Hsu, Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in africa: a clinical validation study, Lancet Digital Health 1 (1) (2019) e35–e44.
[160] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Hannu Uusitalo, Heikki Kälviäinen, Juhani Pietilä, Diaretdb0: evaluation database and methodology for diabetic retinopathy algorithms, Mach. Vis. Pattern Recog. Res. Group Lappeenranta Univ. Technol. Finland 73 (2006) 1–17.
[161] Fei Wang, Lawrence Peter Casalino, Dhruv Khullar, Deep learning in medicine—promise, progress, and challenges, JAMA Int. Med. 179 (3) (2019) 293–294.
[162] David B. Larson, Hugh Harvey, Daniel L. Rubin, Neville Irani, R. Tse Justin, Curtis P. Langlotz, Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations, J. Am. Coll. Radiol. 18 (3) (2021) 413–424.