International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Review on Fundus Image-Based Deep Learning for the Identification and Categorization of Diabetic Retinopathy

Vijayalaxmi Gopu[1], M. Selvi[2]

Submitted: 15/01/2024 **Revised**: 23/02/2024 **Accepted**: 01/03/2024

Abstract: Diabetes-related retinopathy, or diabetic retinopathy, is the most common cause of blindness in the world. Delaying or preventing eyesight loss and impairment calls for prompt diagnosis and treatment. For this reason, several AI-based approaches have been developed for identifying and categorizing diabetic retinopathy in fundus retina images. The application of deep learning techniques in the various stages of the fundus image-based diabetic retinopathy diagnosis pipeline is thoroughly investigated in this review study. From the commonly used datasets in the research community to the preprocessing techniques and how they accelerate and improve model performance, to the creation of deep learning models for diagnosis, grading, and lesion localization, we cover many of the key steps in this pipeline. Some models that have been used in actual clinical practice are also discussed. As a final step, we offer some key takeaways and suggestions for further study.

Keywords: *Deep Learning, retina, diabetic retinopathy, image processing, machine learning.*

1. Introduction

There are currently 463 million individuals with diabetes mellitus, and this figure is expected to climb to 700 million by 2045, making diabetes a major public health concern [1]. Diabetic retinopathy (DR) is the most frequent complication of diabetes affecting the eyes, affecting at least one third of people with diabetes. Any diabetic patient, regardless of how severe their condition is, is at risk for developing DR, which is characterised by increasing vascular disturbances in the retina due to persistent hyperglycemia. It is estimated that some 93 million individuals throughout the world have DR, making it the main cause of blindness among adults of working age.

The increased incidence of diabetes in developing Asian nations like India and China is a major factor in the projected increase in these figures. Neuronal retinal degeneration and clinically undetectable microvascular alterations progress throughout the asymptomatic early stages of diabetic retinopathy. Therefore, it is crucial to test diabetic patients' eyes regularly so that any visual problems can be identified and treated as soon as possible. Early recognition of DR is crucial because the sole method for prevention is the management of risk factors such as hyperglycemia, hyperlipidemia, and hypertension.

In addition, if diagnosed and treated early, proliferative retinopathy and diabetic maculopathy can be prevented in nearly all cases with the use of modern therapies like laser image coagulation. It is now clear that early diagnosis and therapy are crucial in delaying or preventing blindness from diabetic retinopathy. In clinical practice, early diagnosis of DR is dependent on fundus examination, even if it may be predicated on functional alterations in electroretinography (ERG), retinal blood flow, and retinal blood vessel diameter.

One of the most common ways to evaluate the severity of DR is using fundus image, which is a quick, non-invasive, well-tolerated, and easily accessible imaging procedure. Ophthalmologists diagnose and evaluate the severity of diabetic retinopathy by examining fundus images, in which retinal lesions may be seen at high resolution [2]. This has prompted the scientific community to create computer-aided diagnostic tools, which will lessen the burden on human medical professionals in terms of money, time, and effort spent diagnosing DR.

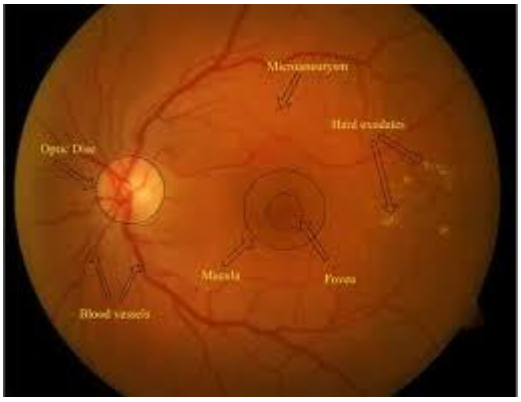

Fig. 1. Fundus image

The possibility to establish Deep Learning (DL) applications for effective DR detection and classification has arisen as a result of recent breakthroughs in AI and the growth in computational resources and capabilities [3]. This article presents and evaluates contemporary DL-based approaches for detecting and classifying DR; that is, those published after 2016. Although some literature reviews on the topic of deep learning's impact on DR have appeared in recent years, they tend to focus on only one or two steps in the data analysis and modelling pipeline (see Fig. 1), such as reporting on the model's performance or on the most

1. *Sathyabama Institute of Science and technology, Chennai, vijayamtech5@gmail.com*
2. *Dept of Computer Science and Engineering, Sathyabama Institute of Science and technology, Chennai, selvi.cse@sathyabama.ac.in*

popular preprocessing techniques. There is no description of the open data sets provided.

A more comprehensive and coordinated effort is required to assess the technological implementations and development in this extremely active research field due to the disjointed nature of the current activities. To accomplish this, we present a new comprehensive description of the analysis pipeline, which includes, among other things, a detailed analysis of the openly accessible datasets, the commonly used preliminary processing conduits.

2. Retinopathy and diabetes

Microaneurysms appear in the retina at an early stage of diabetic retinopathy due to the degeneration and loss of pericytes that causes capillary wall dilatation. Intraretinal hemorrhages develop when a capillary or a microaneurysm bursts. They are all additional pathologies of non-proliferative diabetic retinopathy. Large diameter, tortuous vessels are seen in ischemic regions; these may be IRMAs, as shown by research by Stitt et al. Finally, the occurrence of neovascularization, or the development of new retinal blood vessels in response to ischemia of preexisting vessels, distinguishes non-proliferative from proliferative diabetic retinopathy [4].

Several lesions are shown in Figure 2 on a representative retinal fundus image. Diabetic macular edoema (DME) is the endpoint of any phase of diabetic retinopathy and is the leading cause of blindness. Disc-sized exudates in the macula, thickened retina inside the central fovea, and microaneurysms or haemorrhages inside the central fovea are all associated with edoema. The Experimental Intervention Diabetic Retinopathy Study (ETDRS) grading system is widely recognized as the pinnacle of quality for DR clinical assessment processes; nevertheless, incorporating it into ordinary medical treatment has been challenging. [5].

In an effort to enhance patient screening and communication, several additional scales have been developed. There is currently no standardised worldwide severity measure for diabetic retinopathy despite the development of such streamlined ratings in a few of countries. Thus, the International Clinical Diabetic Retinopathy Disease Severity Scale was proposed by the Global Diabetic Retinopathy Project Group to categorise DR on 5 severity stages.

3. Deep Learning models

Inspired by the architecture of the human brain, Deep Learning (DL) is a family of AI techniques that uses artificial neural networks to learn new skills. Automatically learning the mathematical illustration of the hidden and intrinsic relations in the data is at the heart of deep learning. Since deep learning methods learn relevant features directly from the data, they require significantly less human direction than standard machine learning methods, which are dependent on the development of hand-crafted features, a procedure that may be highly difficult and time-consuming [6].

Also, when data volumes grow, DL approaches perform substantially better than classic ML techniques. In this part, we will quickly go over some fundamental DL ideas. A three-layer artificial neural network (ANN) with one input layer, one hidden layer, and one output layer is the simplest type of neural network. Due to its single hidden layer, such networks are referred to as "Shallow" or "Feed-Forward" Neural Networks. [7].

Unfortunately, imaging data is not a suitable input for these networks since they only take one-dimensional arrays. In contrast to shallow neural networks, which only accept 1D or 2D arrays as input, the notion of Convolutional Neural Networks (CNN) is based on a basic mathematical process called "convolution," which was inspired by human vision. While the output of each neuron in the following layer is calculated using the input of all the neurons in the previous layer in a DNN, this is not the case with a CNN. A CNN, on the other hand, uses filters or kernels to compute convolutions by sliding over a region of the original image to generate a feature map.

Since UNet able to maintain the original image structure, they are superior to conventional CNNs when it comes to semantic segmentation. Specifically, they have a contracting path to capture the important context and an expanding path that is symmetric to allow for exact and accurate segmentation. Furthermore, unlike conventional CNNs, which process the image in multiple passes using a sliding window method, a UNet architecture processes the entire image in a single pass; hence, the name "Fully Convolutional Networks" (FCN) [8]. Last but not least, it performs a segmentation task with considerably less data than conventional CNNs do, which is especially important in the field of medical image analysis, where the amount of data available is far lower than in other areas of computer vision. It's well knowledge that humans can't analyse an entire scene or item all at once, but rather use attention processes to zero down on individual details.

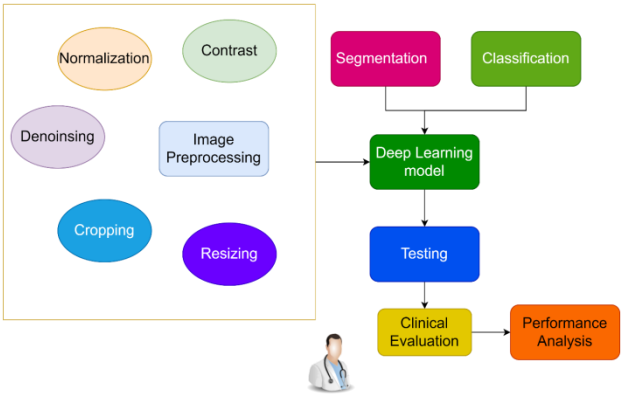

Fig. 2. Process Flow model

However, such a method has not been completely implemented in conventional CNNs. In order to boost the models' efficiency and reliability, several recent research have developed such methods, known as attention modules [36–39]. Finally, the Generative Adversarial Network (GAN) [40] is a significant subclass of convolutional neural networks. The generating network, which creates candidate samples according to the original data distribution, and the discriminator, which tries to tell the difference between the created candidate samples and the genuine data distribution, make up the two halves of a standard GAN. If the generator is trained in this way, it will provide prospective samples that are statistically very similar to the actual data distribution [9].

Training a deep neural network requires a large amount of time, effort, and data. ImageNET [42] is the largest item detection collection in the world, containing over 14 million images of real-

world objects like animals, electronics, food, people, automobiles, and more. However, there are only around 80.000 fundus images in the biggest dataset discussed here. This distinction stems from the fact that unlike images of commonplace items, medical images are notoriously difficult to get on account of the extensive curation, annotation, and legal difficulties that accompany their acquisition [10].

Therefore, it can be quite challenging to train strong and accurate models for medical issues. However, it is feasible to take use of models that have been trained on big datasets, such as ImageNet, by applying the information gained to a different model in a different domain. In order to enhance learning in one activity, Transfer Learning takes previously acquired knowledge and applies it to a new context. Ultimately, the network's ability to recognise low-level elements of the image (i.e. edges, contours, etc.) is improved by the transfer of knowledge from ImageNet to a medical imaging domain. Fine-tuning (i.e. retraining) the model on a fresh task (i.e. new dataset) is required in order to identify DR, however this procedure is far quicker and more precise than training the model from beginning [11].

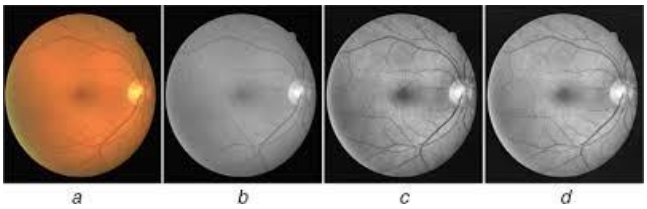

Fig. 3. Prepreocessing of Fundus images

Ensemble Learning, or the use of several approaches, is another key field of research in artificial intelligence. As a potential method to combine data from different modalities, this learning strategy works to lessen the model's generalisation mistake. By combining the results of several different base models trained on the same dataset or a portion of the available data, ensemble learning modelling infrastructure may provide more accurate predictions. If all the models have been trained to do the same job, then it is feasible to utilise a variety of ensemble methods to combine the results. Majority voting, averaging, bagging, stacking, and boosting are all examples of such techniques.

In a classification task, for instance, each classifier generates a result, and the final prediction is based on the result that received the most votes. When solving regression issues, it is common practise to take an average, which may or may not include weighting the individual predictions. Bagging is a method of producing a final prediction by combining the results of many models that were each trained on a small part of the original dataset. In addition, in stacking, a unique meta-model is trained on the output predictions of the various models in order to provide the final prediction. The goal of boosting-based ensemble approaches is to use models that have been trained several times to improve upon the shortcomings of less effective models [12].

Next, the predictive performance of the various models is used to weight the forecasts and provide an average. Multimodal learning, active learning, and the ability to learn while doing other things are also crucial. Figure 3 provides a straightforward example of the methods under discussion. With Multitask Learning, a single data encoding network may be used to make predictions for two distinct tasks. Imaging and non-imaging sources (such as magnetic resonance imaging (MRI), computed tomography (CT), molecular and clinical data)) are often integrated in multimodal learning.

First, data from each source type is analysed by its own specialised model, and then the resulting fused features are utilised to educate a single model. Finally, Active Learning is concerned with training the algorithm on a small marked subset of the data, generating the predictions for the remaining models. The model is then trained using the newly tagged dataset. When the data set is too huge to label manually and labelling should be prioritised, active learning is often utilised.

4. Analysing data on Diabetic retinopathy

A certified expert used the ICDRDSS scale to rate the images. Aravind Eye Hospital collected the data for the Kaggle APTOS 2019 Challenge in rural India so that they may create effective tools for the automated diagnosis of Diabetic Retinopathy and increase their capacity for patient identification. It has 5590 images, making it the third biggest dataset. However, it has certain gaps in coverage, such as the Severe NPDR class having just 193 images [13].

Noise in the data (i.e. artefacts, focus difficulties, being under/overexposed) and the labels are present in the APTOS dataset, just as they are in the Kaggle EyePACS dataset, because they were both gathered in a real-world multicentre scenario. There are 1200 retina fundus images in the Messidor dataset, all of which were acquired in 2005 and 2006 from three different ophthalmology clinics in France. Eight hundred images were taken with pupils dilated and the remaining 400 were taken with normal pupils. Both datasets have high-quality images free of distracting noise, unlike the Kaggle EyePACS dataset.

Each image in the datasets has been assigned a medical diagnostic indicating the degree of Diabetic Retinopathy, but there is no pixel-level information available describing the segmentation of lesions. However, the reliability and generalizability of their grades are diminished since their system deviates from the standard ICDRS procedure. The ophthalmology clinic in Nanded, India, used a Kowa VX - 10 fundus camera to capture these high-quality images, which make up the IDRiD dataset. All subjects had their pupils enlarged before any imagegraphs were taken. Although it has 12522 images and ranks second in size when evaluating the classification challenge, the DDR dataset is still relatively fresh and has not been utilised much. The information was gathered by several experts using a majority vote schema from 2016-2018 from various hospitals and annotated using the ICDRDSS scale. In addition, a sixth grade was added so that low-quality imagegraphs may be placed in their own group [14].

5. Image Processing Approaches

Noise is introduced to the final image when a fundus image is taken with several different kinds of hardware under different kinds of lighting. In majority of the studies we looked at, pre-processing the images was required because it helped minimise heterogeneity, which in turn affected the performance of the classification model, and it helped emphasise some small elements of the images. To begin, in any image processing or analysis pipeline, contrast enhancement is a popular preprocessing method used to emphasise the foreground from the background. Histogram equalisation, a straightforward approach

for boosting contrast in fundus images, does so by increasing the image's global contrast but disregarding local differences [15].

Adaptive Histogram Equalisation is a more sophisticated technique for contrast adjustment that accounts for regional differences in a constrained region of each pixel. Contrast-limited adaptive histogram equalisation (CLAHE) is increasingly widely employed in the scientific community for fundus imaging. In contrast to the original Adaptive Histogram Equalisation technique, CLAHE avoids the problem of excessive contrast amplification in almost constant regions of the image. All of these techniques work by altering the contrast of the image, which makes the retina's fine features more obvious.

To minimise variations in lighting between images and to better show the minor lesions, other researchers have subtracted the local average colour from each pixel and mapped it to 50% grayscale. The image may have noise, thus Non-Local Means Denoising (NLMD) is used to get rid of it. While a more powerful denoising algorithm can successfully remove more noise from an image, it will also cause the image's fine features to blur as a side effect.

Additionally, image intensity normalisation is used to reduce bias and lengthy training durations for the network and to standardise, considering the intensity of its pixels). Using a different colour model or even just the channels has improved the model's performance in addition to contrast augmentation, normalisation, and noise reduction. This approach outperformed the models that effect after being converted to entropy images, as performed by Lin et al.

Additionally, the green channel is typically extracted from the fundus colour image because to its high contrast and plenty of information. Images in the databases could also range in resolution and aspect ratio. There may also be large stretches of blank space in the images. Images may be cropped, rescaled, and resized to a predetermined resolution in order to standardise their size and eliminate such blank spaces [16].

Two distinct cropping methods were used in the tests by Bravo et al. Here, they are cropped so that the circle of the cropped image is surrounded by the retina; in another, the biggest square image inscribed in the retina is cropped. Preprocessing approaches, like as those presented in this section, have been shown to boost performance, notably for fundus images, even though DL has been shown to operate effectively.

Due to the scarcity of high-quality data, data augmentation methods are employed to improve the model's stability and precision. Such methods may include Generative Adversarial Networks for image synthesis, colour and brightness augmentation, and image flipping, rotation, and shifting (translation), resizing, shearing, and flipping. The majority of the examined publications make use of some sort of augmentation technique to boost the available number of images and, thus, speed up the training of the model.

6. Deep Learning Models

These approaches are to categorise DR into five groups (or "classes"), in accordance with the clinical grading system, was published by Pratt et al. To address the class imbalance in the dataset and prevent over-fitting, the authors utilised a class-weighted method to adjust the parameters during backpropagation for each batch.

To improve their ability to forecast NPDR and PDR instances, Islam et al. transformed the quinary (5-class) classification issue into a regression problem. Torre et al. also created a CNN model for doing the categorization; theirs took into account data from both eyes and effectively fused the resulting representations. To get a final receptive field with dimensions as close to the original image as feasible, they suggested utilising tiny convolutions and adjusting the network's architecture [17]. Raju et al. also showed that the DR classification performance was improved while employing smaller (4x4) filters in the Conv2D layers. To try and capture significant lesion markings, which vary in size, in Refs. inception modules have been used to extract features at various resolutions.

While testing their model on two independent datasets, to determine the existence and severity of DR and DME, 54 US-based doctors rated the images anywhere from three to seven times. First, we looked at what happens when we train the model on smaller and smaller pieces of the full dataset, and found that the sweet spot for performance occurs at roughly 60,000 images (with 17,000 referable).

The quality of the image was tested in a second subsampling experiment, and the results showed that the performance improved when the majority voted on the issue. However, Krause et al. investigated the effect of training using an adjudication grading system on the accuracy of the ground-truth labelling. They employed a short dataset in which an adjudication grading methodology was applied to a pre-trained model developed by Gulshan et al. The authors observed that adopting settlement as the milled fact expansion typical resulted in a minor performance gain compared to majority voting.

The detection efficiency of CNNs has also been boosted with the use of attention modules. The attention mechanism and bilinear technique were utilised by Zhao et al. to train a CNN and improve the system's categorization performance in complex regions. To further enhance classification accuracy, the emphasis mappings that Wang et al. created were sent into a Crop-Network, that focused in on the spots with the most focus. A unique architecture was presented by Li et al. that uses attention modules to investigate possible links between diseases in order to concurrently identify DR and DMR.

The original fundus image is employed with the deep learning pipeline established by Lin et al. for DR severity categorization. To mitigate the effect of incomplete lesion annotations, a lesion clustering approach was applied during the detection phase. Every component of a lesion is given a relative priority score, and the Focus Integration System merges these maps with the feature maps generated by the classification algorithm. In order to boost the performance with a system for paying attention from image-level data with annotations, Zhou et al. suggested an ensemble weakly-supervised learning approach [18].

Additionally, high-resolution fundus images would be ideal for training the model. This would allow for the detection of even the smallest lesions. However, this is not possible with deep CNNs because to their high computational complexity and the vanishing/exploding gradient issue. On the other hand, there is a significant loss of detail when images are downsampled directly. Multi-Cell Multi-Task CNN (M2 CNN) is a novel architecture created by Zhou et al. that captures high-resolution details by gradually increasing the network's depth and kernel size in tandem. In the final step, called Multi-Task, a classification score and a regression score are calculated.

Since DR is a chronic condition with no clear inflection points, the authors reformulated the training loss function to be more accurate. The model's performance improves logarithmically with regard to increasing input image resolutions, as determined by Li et al.'s extensive experimentation across a wide range of image resolutions. However, the network complexity grows exponentially with the input image resolution. The best image resolution, given the complexity restrictions, was 896 × 896, which helped the system perform better, especially when classifying the moderate DR scenario correctly, which is dependent on extracting subtle details.

The lack of available data to train the models is a fundamental problem for deep learning in general, and for DL applied to medical imaging in particular. It is feasible to solve this problem by moving expertise from an area with an abundance of data (such as computer vision) to one with a scarcity of data (such as medical imaging). Classification models have been developed using transfer learning in several of the publications we've looked at.

The Kaggle EyePACS dataset was used in a comparison of pre-trained models by Wan et al. When compared to other, more complicated designs, they found that VGGNet's architecture produced the greatest results. Using a previously trained InceptionV3 model from ImageNet, Hagos et al. engaged in transfer learning. The authors adjusted the classifier using a sample of the Kaggle EyePACS dataset that was carefully selected to be balanced. Pre-trained Inception backbone networks (GoogLeNet, InceptionV3, InceptionV4) have been found to produce the greatest results by a number of other researchers. By integrating the benefits of several classifiers, ensemble learning has played a significant role in the development of strong and powerful AI frameworks for DR classification. Given the information gain provided by their complementarity, ensemble learning has been observed to outperform the corresponding solo models. This suggests that due to changes in design or training, the various base models are capable of implicitly learning varying degrees of semantic representations [19].

Two ensemble models were created by Zhang et al.; one was used for illness identification (binary classification), while the other was used for disease grading (quinary classification). Feature extraction in each model was handled by a different set of pre-trained networks, while classification was handled by a bespoke standard dense neural network. The ensemble models achieved a higher sensitivity (98.10%) and specificity (98.56%) than the individual ones. The authors also point out that general performance improved with the 'strength' of the base learner (pre-trained network). The results of a dual ensemble (ensemble of ensembles) were much more impressive than those of a single ensemble.

Three models built on the InceptionV3, ResNet152, and Inception-Resnet-V2 architectures were combined into an ensemble model by Jiang et al. The model was built using an exclusive dataset created in tandem with Beijing Tongren Eye Centre. Sensitivity = 85.57%, Specificity = 90.85%, Accuracy = 88.21%, and AUC = 0.946 were all better than the individual models. Specificity, however, was 91.46 percent higher for InceptionV3. Since, according to the authors, different types of lesions are most effectively recognised at different stages of training, Quellec et al. developed a CNN model and exported it at various stages of the training process. To then predict the severity score of DR, they used ensemble learning (Random Forest Classifier) to integrate the stored models. Sayres et al. assessed the abilities of 10 ophthalmologists under three scenarios: (a) the doctors had access to the raw fundus images; (b) the doctors had access to the DLS grading findings; and (c) the doctors had access to the DLS grading results plus an interpretable heatmap. The heatmaps considered each pixel's impact on the final forecast, which in turn hints about potential lesions [20].

The accuracy of the diagnosis, the rate of subjective confidence in DR rating, and the amount of time spent grading were the major outcomes examined. With model aid, they saw a tendency towards increased accuracy and confidence but also increased grading times. There was a general trend towards higher accuracy and less grading time as readers became more accustomed to using model aid. In addition to this heightened sensitivity, no discernible loss of specificity was noted. Results showed that the grades-only condition benefited all images more than the grades-plus-heatmap condition.

7. Performance Analysis

As was said in the introduction, an experienced ophthalmologist's primary observations are diseased spots on the retina, which play a crucial role in the diagnosis and treatment of diabetic retinopathy. Thus, in this part, we offer details about previously-published deep learning approaches for the automated segmentation of DR-related lesions such exudates, microaneurysms, and hemorrhages.

In contrast to a classification challenge, where the ground truth applies to the whole image, in a segmentation problem, it applies to each individual pixel. Traditional pixel-level measurements like as accuracy, sensitivity, specificity, etc. are deceptive since the image backdrop (the healthy part of the retina) typically dominates the foreground (the real lesions). This is because most of the pixels in the ground truth image represent the healthy retina, and only a small fraction of the pixels represent lesions. Therefore, because the backdrop is primarily matched with itself, a segmentation algorithm's pixel-wise accuracy would continually be virtually flawless, without necessarily properly finding the important lesions. Therefore, the following measures, rather than the standard classification-oriented ones, are best suited to assess a segmentation model's efficacy [21].

The Intersection-over-Union (IoU) is a popular measure for segmentation issues; it is calculated by equating the union of the predicted (P) and ground truth (G) segmentation areas with the overlapping area, IoU = P P G G. It can take on values between 0 and 1, with 1 denoting an exact match and 0 indicating total discord. The evaluation measure is then determined by arithmetically averaging each class's IoUs. The DICE coefficient is another measure; it's calculated by dividing the combined number of pixels in P and G by the area that overlaps both. It has the same worth as the F1score metric. The DICE coefficient, like the IoU, is a numeric value between 0 and 1. Although their absolute values may differ, DICE and IoU are positively associated and hence point in the same direction. This means that while measure A suggests classifier A is superior than classifier B, metric B also suggests this. However, a distinction becomes clear when comparing the relative performance of several classifiers.

The model's efficacy across all decision thresholds is graphically represented by the model curve. A threshold determination for a detected region to be recognised as genuine or not is required for FROC, unlike ROC curves. One possible threshold for a positive

detection is an overlap of 50% between the annotated and detected areas [22].

They can also be used to evaluate a segmentation method. However, as none of these criteria are referenced or employed in the publications under evaluation, they will not be discussed at length here. In order to concurrently segment all four DR-related lesions in a fundus image (Soft/Hard Exudates, Haemorrhages, and Microaneurysms), Guo et al. suggested L-Seg network. Their approach generates four distinct segmentation maps, one for each class of lesion. To get around class- and loss-imbalance difficulties, they also suggest a multi-channel bin loss function that uses all four outputs together. They used a weighted fusion module to combine all this data and successfully analyse complicated lesions, as well as numerous feature maps of the network to encompass multi-scale analysis and manage lesions of varying sizes.

The authors describe an "interestingness" score that ranks the remaining unlabeled patches according to how much information they contain about the target lesion. They start by using the labelled portion of the dataset to train the network. Even if not all images were input to the model, training is terminated once convergence is achieved. This guarantees that only the most informative images are used. For the purpose of patch generation, two distinct partitioning strategies were studied. The first method included taking a 48x48 pixel patch surrounding a randomly chosen pixel within an exudate lesion. This method guaranteed that the chosen patches had exudates, however exudates from different patches may overlap.

The second method included removing little sections of the image at a time. While no two patches were next to one another, only a small fraction of those patches really had an exudate. When the model was trained using a combination of 75% patches from the first method and 25% patches from the second, the F1score was 92.8%. In order to show where the lesions are to eliminate unnecessary borders and messy pixels surrounding the segmented signs. They created three probabilistic output maps, one for each lesion, however it's possible that some pixels correspond to more than one lesion.

To get around this problem, they decided to give each pixel an assignment based on which of the three possible lesions it was more likely to belong to. To better visualise potential lesion regions that contributed significantly to the networks' primary goal is to categorise the image according on its severity. On the other hand, such a rough estimate of the lesion regions might be viewed as an approximation of the area the network is concentrating on while making its choice [23].

To be more specific, during training, the networks (regardless of design) picked up on hard and soft exudates more sooner than they did on haemorrhages and then microaneurysms. It is common for high-quality images to have been collected under controlled, non-standard settings for many datasets, such as Messidor, IDRiD, etc. It follows that it is possible to claim that algorithms trained on such datasets may not fare well in more typical actual scenarios, where the images may not be precisely similar and where environmental and hardware aspects will change. Although the Kaggle EyePACS and APTOS datasets solve these difficulties and closely mirror a real-world scenario, the noise which is there due to those fluctuations makes it exceedingly difficult for the algorithms to precisely and efficiently complete the analysis.

Nonetheless, by using such low-quality images as a representation of the actual data, one might create strong algorithms with potential for use in clinical practise. bias and provide reliable and accurate ground truth data for certain kinds of data. To improve the reliability of the ground truth and hence the precision of the model, Gulshan et al. also suggested having numerous experts independently assess the obtained data.

To that end, it is important to create a common benchmark against which all graders may compare their work. In contrast to a majority decision methodology, an adjudication grading standard was found to be more accurate in the later trial in spotting artefacts and missing microaneurysms. Both the training process and the model's final results might be negatively impacted by the data's poor image quality. Images with poor contrast or blurriness might obscure the earliest, most subtle indicators of retinopathy.

In their diagnostic workflow, Rakhlin et al. suggested a quality evaluation module to remove ungradable images. After that, an ophthalmologist is brought in to have a look at the imagegraphs. Even Jiang et al. omitted the blurry, low-resolution images from their dataset of choice. Using the ICDR grading scale approach as the basis for the first five classes, Li et al. included the quality evaluation module into the deep learning model and posed the classification as a six-class scenario grading problem.

Tan et al. analysed data gathered from 11 distinct clinical locations using different types of fundus cameras. They used an Ascore to evaluate the image's quality during the normalisation process. The quantity of grey pixels that arose during normalisation of the image's dark regions served as the basis for its calculation. Images with such dark regions were removed from both the training and testing sets since their illumination revealed no useful information about the retina's anatomy.

However, Quellec et al. state that the performance of their ensemble model was not significantly impacted by image quality. But the model's performance may be enhanced, or at least made easier to train, by enhancing the uniformity and consistency among the data. This can be done by either manipulating the camera settings and ambient variables during the collection, or by removing low-quality images [24].

One of the many processes required to create reliable and effective AI models is the creation of massive training and assessment datasets. However, the aforementioned datasets typically lack sufficient data or have class imbalance. Fortunately, this problem may be circumvented in a number of ways, including the use of augmentation techniques, the generation of synthetic data using GANs, transfer learning to draw on the expertise of models trained on big datasets like ImageNET, and so on.

To verify the model's reproducibility, it's also crucial to collect more demographically diverse data. It was also suggested by Gargeya et al. that additional patient information be incorporated, such as the patient's genetic makeup, the length of time they've had diabetes, their haemoglobin A1C level, and any other clinical data that may affect their risk for getting retinopathy. Incorporating data on explicit lesion aspects into the classification models could be useful as well. By doing so, the AI model may discover surprising associations between previously unrelated data, providing new insight into the causes of DR and improving diagnosis accuracy [25].

8. Research Directions

The healthcare industry might benefit from the application of AI and, more specifically, Deep Learning (DL)-based methodologies. However, a number of significant barriers must be removed before AI can be widely used in hospitals and other healthcare facilities. The traditional methods which are used to test the performance of the approach, i.e. accuracy metrics, are given as only one of several ways that are key components towards the acceptance of AI models through regulatory procedures. Despite improving the effectiveness of such research, the transition from traditional [26-27] machine learning approaches to deep learning ones has been accompanied by a lack of explainability and transparency. However, their interpretability is a critical factor impacting their adoption and incorporation into clinical practise. The model's decision process, which should ideally include justifications for its predictions (e.g., whether those projections have been chosen and what alternatives were examined), must be grasped by the clinical operator. In an effort to quantify the relative contribution of each pixel across all of the network's layers in making a DR prediction, a number of researchers have created evidence heatmaps. These representations let doctors see if the model is basing its forecast on important clinical aspects, such exudates, microaneurysms, and haemorrhages in the case of DR. The models' robustness and dependability are other important considerations that must be taken into account before they can be used in a clinical setting. In the end, these phrases reflect to the requirement that the models continually function properly despite the many unexpected variances present in the clinical environment, such as differences in data obtained from different centres or devices manufactured by different manufacturers.

9. Conclusion

The increasing impairment and eventual loss of vision caused by diabetic retinopathy is a devastating consequence of diabetes mellitus. In order to prevent further decline and retinal damage, early diagnosis and treatment is crucial. Over the past several years, there has been a surge in enthusiasm for using DL systems for diagnosing diabetic retinopathy, and when these systems mature and are eventually integrated into clinical practise, they will help doctors better care for their patients. This article summarises what is known so far about using deep learning to detect diabetic retinopathy. The field of ophthalmology has benefited greatly from the advent of deep learning, but there is still room for development in terms of performance, interpretability, and reliability.

References

[1] Fourcade, A., & Khonsari, R. H. (2019). Deep learning in medical image analysis: A third eye for doctors. *Journal of stomatology, oral and maxillofacial surgery*, *120*(4), 279-288.

[2] Ting, D. S. W., Cheung, C. Y. L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., ... & Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, *318*(22), 2211-2223.

[3] Abbas, Q. (2017). Glaucoma-deep: detection of glaucoma eye disease on retinal fundus images using deep learning. *International Journal of Advanced Computer Science and Applications*, *8*(6).

[4] Sarki, R., Ahmed, K., Wang, H., & Zhang, Y. (2020). Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. *IEEE access*, *8*, 151133-151149.

[5] Nazir, T., Nawaz, M., Rashid, J., Mahum, R., Masood, M., Mehmood, A., ... & Hussain, A. (2021). Detection of diabetic eye disease from retinal images using a deep learning based CenterNet model. *Sensors*, *21*(16), 5283.

[6] Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., ... & Campilho, A. (2020). DR| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, *63*, 101715.

[7] Butt, M. M., Iskandar, D. A., Abdelhamid, S. E., Latif, G., & Alghazo, R. (2022). Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features. *Diagnostics*, *12*(7), 1607.

[8] Butt, M. M., Iskandar, D. A., Abdelhamid, S. E., Latif, G., & Alghazo, R. (2022). Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features. *Diagnostics*, *12*(7), 1607.

[9] Stember, J. N., Celik, H., Gutman, D., Swinburne, N., Young, R., Eskreis-Winkler, S., ... & Bagci, U. (2020). Integrating eye tracking and speech recognition accurately annotates MR brain images for deep learning: proof of principle. *Radiology: Artificial Intelligence*, *3*(1), e200047.

[10] Triwijoyo, B. K., Sabarguna, B. S., Budiharto, W., & Abdurachman, E. (2020). Deep learning approach for classification of eye diseases based on color fundus images. In *Diabetes and Fundus OCT* (pp. 25-57). Elsevier.

[11] Reiher, L., Lampe, B., & Eckstein, L. (2020, September). A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-7). IEEE.

[12] Stember, J. N., Celik, H., Krupinski, E., Chang, P. D., Mutasa, S., Wood, B. J., ... & Bagci, U. (2019). Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of digital imaging*, *32*, 597-604.

[13] Nuzzi, R., Boscia, G., Marolo, P., & Ricardi, F. (2021). The impact of artificial intelligence and deep learning in eye diseases: a review. *Frontiers in Medicine*, *8*, 710329.

[14] Dos Santos, V. A., Schmetterer, L., Stegmann, H., Pfister, M., Messner, A., Schmidinger, G., ... & Werkmeister, R. M. (2019). CorneaNet: fast segmentation of cornea OCT scans of healthy and keratoconic eyes using deep learning. *Biomedical optics express*, *10*(2), 622-641.

[15] Yoshimura, Y., Cai, B., Wang, Z., & Ratti, C. (2019). Deep learning architect: classification for architectural design through the eye of artificial intelligence. *Computational Urban Planning and Management for Smart Cities 16*, 249-265.

[16] Nazir, T., Irtaza, A., Javed, A., Malik, H., Hussain, D., & Naqvi, R. A. (2020). Retinal image analysis for diabetes-based eye disease detection using deep learning. Applied Sciences, 10(18), 6185.

[17] Chea, N., & Nam, Y. (2021). Classification of fundus images based on deep learning for detecting eye diseases.

[18] Jain, L., Murthy, H. S., Patel, C., & Bansal, D. (2018, December). Retinal eye disease detection using deep learning. In *2018 Fourteenth International Conference on Information Processing (ICINPRO)* (pp. 1-6). IEEE.

[19] Sarki, R., Ahmed, K., Wang, H., & Zhang, Y. (2020). Automated detection of mild and multi-class diabetic eye diseases using deep learning. *Health Information Science and Systems*, *8*(1), 32.

[20] Babenko, B., Traynis, I., Chen, C., Singh, P., Uddin, A., Cuadros, J., ... & Liu, Y. (2023). A deep learning model for novel systemic biomarkers in photographs of the external eye: a retrospective study. *The Lancet Digital Health*.

[21] Nojiri, N., Kong, X., Meng, L., & Shimakawa, H. (2019). Discussion on machine learning and deep learning based makeup considered eye status recognition for driver drowsiness. *Procedia computer science*, *147*, 264-270.

[22] Ganguly, B., Biswas, S., Ghosh, S., Maiti, S., & Bodhak, S. (2019, January). A deep learning framework for eye melanoma detection employing convolutional neural network. In *2019 international conference on computer, electrical & communication engineering (ICCECE)* (pp. 1-4). IEEE.

[23] Mazzeo, P. L., D'Amico, D., Spagnolo, P., & Distante, C. (2021, September). Deep learning based eye gaze estimation and prediction. In *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)* (pp. 1-6). IEEE.

[24] Poddar, D., Nagori, S., Mathew, M., Maji, D., & Garud, H. (2019, July). Deep learning based parking spot detection and classification in fish-eye images. In *2019 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1-5). IEEE.

[25] Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., ... & Weber, B. H. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, *125*(9), 1410-1420.

[26] Saravanan, P., Aparna Pandey, Kapil Joshi, Ruchika Rondon, Jonnadula Narasimharao, and Afsha Akkalkot Imran. "Using machine learning principles, the classification method for face spoof detection in artificial neural networks." In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 2784-2788. IEEE, 2023.

[27] Y. Aparna, G. Somasekhar, N. Bhaskar, K. S. Raju, G. Divya and K. R. Madhavi, "Analytical Approach for Soil and Land Classification Using Image Processing with Deep Learning," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/INOCON57975.2023.10101169.