Medical & Biological Engineering & Computing (2022) 60:633–642
https://doi.org/10.1007/s11517-021-02487-8

REVIEW ARTICLE

The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review

Wei Xiang Lim[1] · ZhiYuan Chen[1] · Amr Ahmed[1]

Received: 5 January 2021 / Accepted: 14 November 2021 / Published online: 27 January 2022

Abstract
Diabetic retinopathy (DR) is a chronic eye condition that is rapidly growing due to the prevalence of diabetes. There are challenges such as the dearth of ophthalmologists, healthcare resources, and facilities that are unable to provide patients with appropriate eye screening services. As a result, deep learning (DL) has the potential to play a critical role as a powerful automated diagnostic tool in the field of ophthalmology, particularly in the early detection of DR when compared to traditional detection techniques. The DL models are known as black boxes, despite the fact that they are widely adopted. They make no attempt to explain how the model learns representations or why it makes a particular prediction. Due to the black box design architecture, DL methods make it difficult for intended end-users like ophthalmologists to grasp how the models function, preventing model acceptance for clinical usage. Recently, several studies on the interpretability of DL methods used in DR-related tasks such as DR classification and segmentation have been published. The goal of this paper is to provide a detailed overview of interpretability strategies used in DR-related tasks. This paper also includes the authors' insights and future directions in the field of DR to help the research community overcome research problems.

Keywords Deep learning · Interpretability · Diabetic retinopathy · Review

1 Introduction

DR is an eye disorder that develops over time as a result of diabetes. It is the most common cause of retinal visual impairment. It causes capillary leakage and obstruction in the retinal capillaries. DR is a form of silent disease that the patients become aware of when they experience vision loss. With an increase in diabetes patients' life expectancy, the prevalence of DR has grown. Diabetes is one of the most pressing issues in today's field of healthcare [1]. Over the next 25 years, the WHO predicts that the number of individuals with diabetes would rise from 130 to 350 million [1].

The treatment for DR is becoming more complex. [2]. Based on the presence of retinal lesions and retinal vessel variations, DR is divided into two categories: non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [3]. One of the most important and efficient ways to track the progression of DR is through eye screening. It helps professionals to recognize the early signs and symptoms of DR, as well as treat and prevent the disease's progression and vision loss. However, providing patients with eye screening services is a challenge. The main problem is the scarcity of experienced ophthalmologists, as well as healthcare resources and facilities [4]. Because the doctor-to-patient ratio is so skewed, manually analyzing fundus photographs is becoming increasingly difficult [5]. Furthermore, because this form of image-based diagnosis necessitates daily practice, training new staff for it is a lengthy procedure [6].

DL, a subfield of machine learning, is concerned with algorithms inspired by the structure and functions of the brain and could be used as a diagnostic tool for DR. For DR classification, DL approaches, particularly convolutional neural networks (CNNs), have lately been applied with promising results. These DL models adaptively develop the best representation using raw image data as input rather than depending on human feature extraction. Mateen et al. [7] adopted a VGG-19 CNN model to detect the DR lesions in fundus images with an accuracy of 98.34%. In another study, Gulshan et al. [8] developed a CNN model to detect DR and

✉ Wei Xiang Lim
hcxwl1@nottingham.edu.my

[1] Faculty of Science and Engineering, School of Computer Science, University of Nottingham, Jalan Broga, 43500 Semenyih Selangor Darul Ehsan, Malaysia

diabetic macular edema (DME) using fundus images. Their diagnostic result was 96.1% sensitivity and 93.9% specificity, respectively.

DL models, on the other hand, are known to be "black boxes" [9–11]. CNN models make no attempt to explain what representations have been learned or why a certain prediction is generated. One fundamental shortcoming of CNN models, according to Gulshan et al. [8], is that the neural network only receives the input image and associated true class. No precise definitions of the characteristics that would explain the medical diagnostic exist. As a result, after grading, the prediction of a DR diagnosis can be phrased as a classification issue; thus, the diagnostic process is a "black box." This lack of interpretability may make it difficult for target end-users, such as ophthalmologists, to comprehend how the models function, preventing model acceptance for clinical application. As a result, DL-based diagnosis systems have an impact on not only information on ethics but also responsibility, safety, and industrial liability because of their "black box" design architecture and absence of human verification [12].

There is not much debate about the interpretability of these CNN models: where did the networks look for discriminative characteristics when diagnosing DR? While classification accuracy is critical in automated diagnosis activities, understanding the reasoning behind the computer-assisted conclusion has become increasingly important and valued. As a result, the goal of this review paper is to examine the current state of interpretability strategies used in DR-related tasks. This review paper's aim is to identify the benefits and drawbacks of DL interpretability techniques in DR-related activities. The purpose of this review paper is to inform readers about the limitations of interpretability methodologies so that future research areas to improve the interpretability of DL models can be presented.

The remainder of the paper is laid out as follows. The selection criteria for the reviewed articles are explained in Sect. 2. The overall DL pipeline adopted in the reviewed literature is briefly explained in Sect. 3. The available state-of-the-art interpretability techniques for DR diagnosis are highlighted in Sect. 4. The merits and drawbacks of the identified interpretability strategies will be explored in Sect. 5. Finally, in Sect. 6, we conclude the review study by summarizing the most important findings from the state-of-the-art analysis and discussing the future research directions.

2 Articles selection criteria

Figure 1 depicts the overall selection criteria for the reviewed publications in this paper. Despite the fact that this review paper is not a systematic review paper, the overall procedure is very similar to that of a systematic review paper.

To begin, a keyword search was conducted in two academic databases with our review goal in mind. The university's e-library database and Google Scholar are the two academic databases. The primary review objective was chosen using five filters. The following five filters were used to pick articles: (i) target keywords, (ii) publication year, (iii) article title, abstract, and keyword screening for article selection, (iv) cross-checking references of selected publications, and (v) final quality rating of the selected article. The target keywords were searched using the "AND" Boolean operator and included "deep learning." "fundus images," "diabetic retinopathy," "diabetic eye disease," "diabetic retinal disease," "deep learning interpretability," "deep learning explainability," "post hoc interpretability," "explainable artificial intelligence," and "explainable AI". Articles published between 2017 and 2021 were considered eligible for this study due to the rapid advancement in the field. After the selection process, duplications of articles were filtered out. In addition, we also studied the bibliography and citation of the selected articles. Finally, a quality assessment of filtered articles was carried out.

3 Overview of deep learning pipeline in DR-related tasks

The DL pipeline identified in the selected reviewed literature can be generalized as depicted in Fig. 2. To improve the image quality and features, a series of image preparation processes is first established. Geometric and photometric perturbations are two types of picture preprocessing methods. Image resizing, cropping, and flipping are examples of geometric perturbations, whereas modification of image color, brightness, and contrast are examples of photometric perturbations. We discovered that not all authors went through the image preprocessing phase such as [13–15]. Authors from [16–18] on the other hand, preprocess the fundus photos before starting model training. Some other

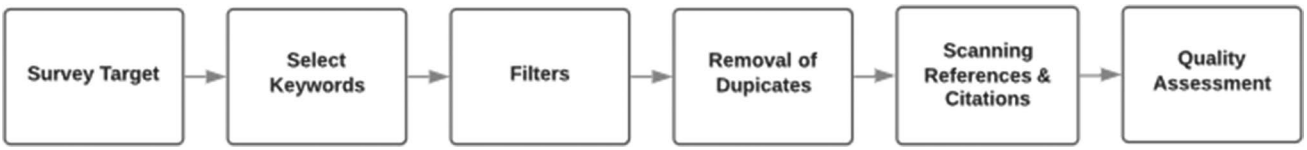

Fig. 1 Overview of selection criteria of literature

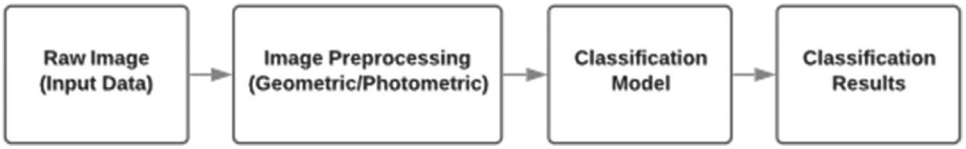

Fig. 2 Overview of deep learning pipeline in DR-related tasks

authors [19–23] partially utilized geometric or photometric perturbation to preprocess the fundus images.

The classification model is trained using the preprocessed input photos. Different authors have used different CNN models to train their networks. Instead of adopting state-of-the-art models, some writers prefer to develop their own CNN model to integrate flexibility [15, 19–21]. Some people choose to use cutting-edge models like Inception-v3, Inception-v4, and DenseNet [13, 14, 22, 24]. Others choose to use data science competition-winning models [16, 18], while others prefer ensemble models [17]. Table 1 summarizes the image preprocessing strategies and the adopted model architectures from all selected papers.

4 Deep learning interpretability techniques for DR

Interpretability approaches for DL are tools that aid in understanding the reasoning behind a neural network's predictions. In recent years, a number of strategies have been proposed. The importance of DL model interpretability and the taxonomy of interpretability methodologies have been carefully reviewed in a number of published review articles [9–11]. In this section, we will go through the many sorts of interpretability strategies used in DR-related tasks. The identified interpretability techniques that are adopted in DR-related tasks include occlusion, sensitivity analysis, class activation map (CAM), gradient-weighted class activation map (Grad-CAM), layer-wise relevance propagation (LRP), and integrated gradient (IG).

A. Occlusion Occlusion is a perturbation technique. The perturbation technique involves removing, masking, or changing a set of input features, then executing a second forward pass on the input features and assessing the effect on the output [25]. Given a scenario, when an image is correctly classified; a natural concern arises as to whether the model is accurately recognizing the object in the image. The occlusion technique answers this question by systematically replacing different regions of the image with a gray square and monitoring the model's output. When the gray square covers the object in the image, the probability of the correct class drops significantly. Zeiler and Fergus [26] reported that the significant drop in probability indicates that the model is correctly detecting the object in the image.

In DR-related tasks, Grassmann et al. [17] adopted the occlusion technique by randomly masking 10,000 100×100 pixel fields in each 512×512 fundus image. The occlusion technique was used to successfully occlude

Table 1 Comparison table of recently published DR-related papers that adopt interpretability techniques

Author (year)	Image pre-processing		CNN architecture	Interpretability techniques
	Geometric perturbation	Photometric perturbation		
Kermany et al. [13]	✗	✗	Inception-v3	Occlusion
Grassmann et al. [17]	✓	✓	Ensemble Network	Occlusion
Kumar et al. [21]	✓	✗	CNN	CAM
Tu et al. [15]	✗	✗	CNN	CAM
Wang and Yang [23]	✓	✗	Net-4, Net-5	CAM
Gargeya and Leng [20]	✓	✗	CNN	CAM
Gondal et al. [16]	✓	✓	o_O network (Antony and Brüggemann 2015)	CAM
Jiang et al. [24]	✓	✓	RestNet50	Grad-CAM
Pratt et al. [14]	✗	✗	DenseNet-121	Sensitivity analysis, CAM
Sayres et al. [22]	✓	✗	Inception-v4	Integrated gradient
Quellec et al. [18]	✓	✓	o_O network (Antony and Brüggemann 2015), AlexNet	Layer-wise relevance propagation
de La Torre et al. [19]	✓	✗	CNN	Layer-wise relevance propagation

essential age-related macular degeneration (AMD) characteristics and allow assessment of their relevance. They noted a considerable decline in prediction scores as the slider went through the occluded zone. As indicated by the significant drop in prediction scores, the occluded region was identified as being a critical component for their CNN model to label a particular fundus image for a specific class.

B. Sensitivity analysis Sensitivity analysis (SA) is one of the first methods to be adopted in the DL domain [27]. Attributions explain individual predictions by assigning weights to each input characteristic based on how much it influences the outcome positively or negatively. They are constructed by taking the absolute value of the partial derivative of the target class score S_c with respect to the inputs x_i:

$$R_i^c(x) = |\frac{\partial S_c(x)}{\partial x_i}|$$

The SA works by taking the absolute value of the gradient, which indicates which input features (e.g., input pixels or tabular data) may be perturbed the least in order for the target output to change the most while eliminating any information regarding the change's direction [27].

In a DR-related task, Pratt et al. [14] employed SA to provide insight into the features that were detected utilizing ground truth and input fundus images. They claimed that by employing the SA method, they were able to pinpoint pixels (features) that aided in diagnosis. Numerous types of lesions, such as hemorrhages and microaneurysms, can be seen all over the vascular structure. The SA can recognize lesions in instances classified as proliferative. Although SA is a good performer, it is affected by noisy gradient, thus noisy visualization.

C. Class activation map The CAM technique inspects a new input image and determines which parts or pixels of the image have contributed most to the model's final output. Formally, CAM interprets a CNN by linearly combining activation maps from the last fully-connected layer, together with its last layer's fully connected weights that correspond to a target class [28]. In other words, a class activation map of a particular class indicates the discriminative image regions used by the CNN to identify that class [29].

Formally, let $I(x, y)$ be a given image. In the last convolution layer, the activation of a node k is $f_k(x, y)$. Then, for unit k, the result of performing global average pooling is $F_k = \sum_{(x,y)} f_k(x, y)$. Thus, the softmax for a given class c is $S_c = \sum_k w_k^c F_k$ where w_k^c is the weight corresponding to class c for unit k. Next, $P_c = \frac{exp^{S_c}}{\sum_{c=0} exp^{S_c}}$ is then generated as the softmax output.

By connecting $F_k = \sum_{(x,y)} f_k(x, y)$ into the class score, S_c, CAM can be defined as below:

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) \tag{1}$$

$$S_c = \sum_{x,y} \sum_k w_k^c f_k(x, y) \tag{2}$$

$$CAM_c = \sum_{x,y} w_k^c f_k(x, y) \tag{3}$$

Unique network architecture is required for the CAM approach, which includes a global average pooling layer after the last convolutional layer and a linear (dense) layer after that. As a result, existing networks that lack this structure cannot use the CAM approach.

Pratt et al. [14] used the CAM technique (Fig. 3) to visualize the learned features and their locations in order to figure out how the CNN model arrives at its prediction and how that relates to manual feature-based grading. The CAM visualization directly relates to the prediction score of the class. Similarly, Gargeya and Leng [20] adopted the CAM approach to emphasize the locations that were critical to their model's prediction. The highlighted regions, according to the authors, represent essential aspects that ophthalmologists employ to reach a diagnosis, and the highlighted region supports their model's domain-guided learning method. Other notable papers have also adopted the CAM method to visualize the regions that are important for the CNN model's prediction [16, 20, 21, 23].

Although the CAM method can qualitatively highlight the learned features and their respective location in an input image, there is no quantitative metric to evaluate the highlighted learned features that are true to the prediction.

D. Gradient-weighted class activation map The Grad-CAM technique is an extension of the CAM technique. The Grad-CAM technique differs from the CAM technique in that it integrates gradient information. It employs the gradients of any target concept (for example, logits for the "cat" category if the input data is an image), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. The Grad-CAM follows the same procedure as CAM, with the addition of computing the weighted sum of the activations and then upsampling the result to the image size before plotting the original image with the visualization (heatmap) [30].

Formally, assumed that the output map of the last convolution layer was denoted as A^k, where k is the number of

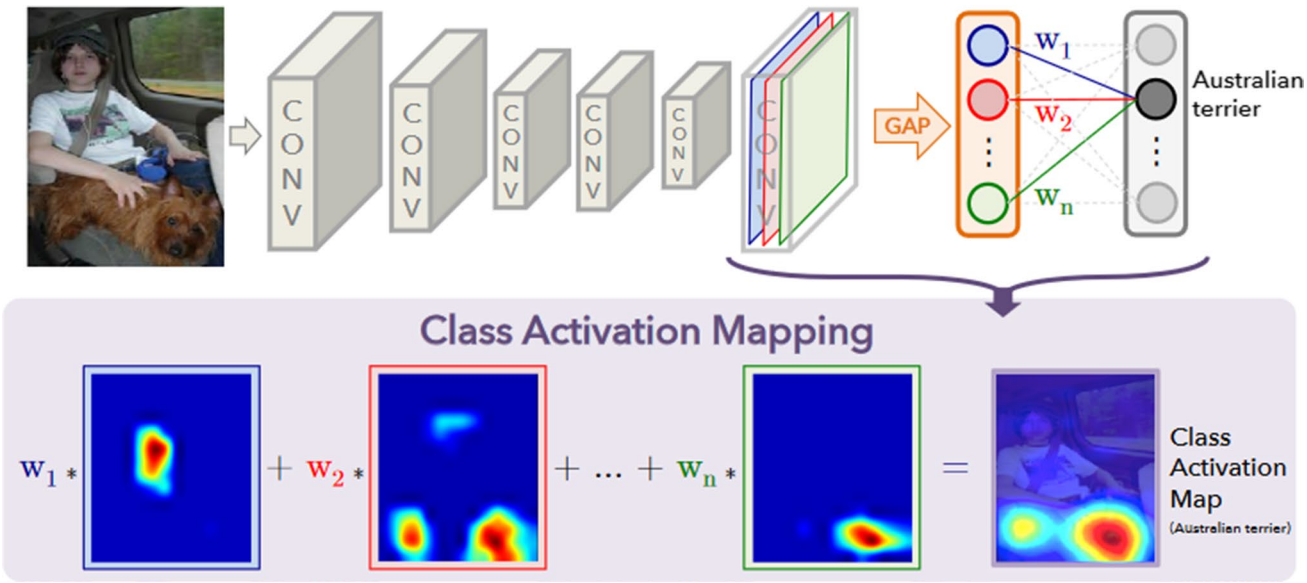

Fig. 3 An illustration of the class activation map [29]

these output maps. The final Grad-CAM can be calculated as follows:

$$w_k^c = \frac{1}{Z}\sum_{i=1}^{W}\sum_{j=1}^{H}\frac{\partial y^c}{\partial A_{ij}^k}$$

$$I_{Grad-CAM}^c = \mathrm{ReLU}\sum_{k=1}^{K}\left(w_k^c \bullet A^k\right)$$

where y^c signifies the class c score before the softmax and A^k is $W \times H$ in size. w_k^c is derived as the map A^k for class c, using a differential operation of y^c with regard to A^k, and Z is the normalization factor. An activation function of the rectified linear unit (ReLU) was implemented after completing a weighted summation of map A^k.

Recently, Jiang et al. [24] proposed a novel approach to simultaneously complete DR classification and detection tasks based on multi-label and Grad-CAM. Different DR lesions were treated as different labels for DR fundus pictures, removing the requirement for annotation experts to provide the exact location of lesions on the fundus image, which not only saved time but also reduced incorrect or omitted to the label of lesions. Grad-CAM was also used to show the precise location of lesions in order to explain the categorization results.

E. Layer-wise relevance propagation LRP is an interpretability technique that uses input variables to explain individual DL model predictions. It assigns each of the input variables a score (relevance) for each input and the model's prediction, indicating how much they contributed to the prediction. In other words, the LRP technique requires modification to the back-propagation rules so that a back-propagated signal is weighted by the activations and weights of convolution layers. Furthermore, the LRP technique employs heuristic propagation rules that apply to each layer of a network to reverse-propagate the prediction via the network [31].

The LRP technique embodies a core principle which is total relevance. For example, the activation strength of an output node for a given class is conserved per layer; each of the nodes in layer l that contributed to the activation of a node j in the subsequent layer $l+1$ is given a certain share of that node's relevance R_{l+1}^j. Overall, the relevance of all nodes i contributing to neuron j in layer l must add up to R_{l+1}^j, preserving the overall relevance per layer:

$$\sum_i R_{l,l+1}^{i\to j} = R_{l+1}^j$$

de La Torre et al. [19] applied LRP on the last layer in order to find the optimal number of components that maximize the classification capabilities using features with smaller dimensions. de La Torre et al. [19] modified LRP in such a way that reducing it to only three components is possible to achieve almost 99% of the evaluation metric. Such value experimentally proves that the principle component analysis (PCA)—a strategy for reducing the dimensionality of a large dataset by converting a larger collection of variables into a smaller set that retains the majority of the information in the larger set, has been able to extract all the crucial elements required for the classification.

F. Integrated gradient IG is an attribution method that was proposed by Sundararajan et al. [32]. It is an

interpretability technique that visualizes the importance of input features that contribute to the model's prediction. The attribution of each input to the output neuron is calculated by looking at the change of this neuron when the input shifts from the baseline input to the target input of interest. Formally, let x represent the input of interest, x' represent the baseline input, and $F(\bullet)$ represent the neuron of interest. Then, the IG for the i^{th} input variable x, can be approximated by

$$IntegreatedGrads_i^{approx}(x) = \frac{x_i - x_i'}{m} \sum_{k=l}^{m} \frac{\partial F(\bar{x})}{\partial \bar{x}_i}, \bar{x} = x' + \frac{k}{m}(x - x')$$

The number of steps utilized to approximate the integral is denoted by m. The baseline x commonly used as a zero-based input sequence.

Quellec et al. [18] conducted a study using DL to detect lesions or other biomarkers of DR and provide interpretation at the image and lesion level using visualization methods. Their model achieved 0.97 AUC with 94 and 98% sensitivity and specificity, respectively, on fivefold cross-validation using their local dataset. They stated that heat maps produced tend to contain false artifacts caused by the CNN architecture. Thus, the authors proposed reducing those unwanted artifacts through joint optimization of the CNN predictions and the produced heat maps. Specifically, they reduced those unwanted artifacts, modifying the sensitivity analysis algorithm by forcing local changes to preserve the hue. By doing this, the modified sensitivity analysis can focus on pattern enhancement. Quellec et al. [18] also stated that it tends to increase the influence of other lesions near the correct lesions. This limitation was due to the downsampling effect, which occurred in pooling or convolution layers with strides greater than one. To mitigate this problem, the authors used a brute-force solution for reducing these artifacts by introducing additional regularization to the loss function.

In summary, the aforementioned interpretability techniques have their own merits and drawbacks. Table 2 summarizes the limitations of each interpretability technique. While these aforementioned interpretability techniques were applied in DR-related tasks, they also can be adopted in other biomedical tasks. In the midst of the COVID-19 pandemic, [33] used machine learning to guide diagnosis from lung ultrasonography. The authors adopted the CAM approach to identify patterns that influenced the model's judgment on whether bacterial or viral pneumonia was present in the lungs. [34] used the IG technique to illustrate model attribution in the biomedical field of a breast cancer diagnosis. They claimed that models with more segmentation channels were better at focusing on specific parts of the image containing abnormal cells.

5 Discussion

Table 1 summarizes the recently published DR-related papers with the adoption of deep learning interpretability techniques. The table includes a number of factors that we believe are important to diagnose DR. The factors are the use of image pre-processing, the type of CNN architecture, and importantly the adopted interpretability techniques. Table 2 summarizes the limitations of the adopted interpretability techniques in DR-related tasks.

Table 2 The limitations of interpretability techniques adopted in recently published DR-related papers

Author (Year)	Limitation
Kermany et al. [13]	■ Diabetic macular edema and choroidal neovascularization did not highlight a clear point of interest ■ Requires additional mode training with occluded fundus image to monitor the classification score of a particular class
Grassmann et al. [17]	■ The accuracy of identifying the disease can be improved by including additional disease features ■ Requires additional mode training with occluded fundus image to monitor the classification score of a particular class
Kumar et al. [21]	■ Highlight on the neovascularization of the optic disc is absent ■ Visualization (CAM) on lesions areas may not be accurate as there is no pixel-level ground truth presented
Tu et al. [15]	■ Minority of the important lesion areas are highlighted as low impact after lesions regularization
Wang and Yang [23]	■ Fails to highlight the correct lesions that correspond to a class
Gargeya and Leng [20]	■ Visualization (CAM) on lesions areas may not be accurate as there is no pixel-level ground truth presented
Gondal et al. [16]	■ Lesion, specifically red small dots did not detect and highlight accurately
Jiang et al. [24]	■ Visualization (CAM) on lesions areas may not be accurate as there is no pixel-level ground truth presented
Pratt et al. [14]	■ Visualization (CAM) on lesions areas may not be accurate as there is no pixel-level ground truth presented
Sayres et al. [22]	■ Visualizing lesions in misclassified cases may cause over-diagnosis
Quellec et al. [18]	■ Poor quality of visualization of lesions (hemorrhages and microaneurysms) in AlexNet
de La Torre et al. [19]	■ Noisy visualization of lesions

6 Image manipulation

Fundus images play an essential role in diagnosing DR. Ophthalmologists analyze fundus images by identifying retinal lesions or the variations of vessel structure [35]. Therefore, the quality of fundus images is essential [36]. Fundus images may contain artifacts such as noise and varying contrast levels caused by the environment settings and use different models of cameras. Pre-processing fundus images help to overcome these issues and allows the retinal lesions to become more prominent, and make it easier for CNNs to extract the features in the fundus images.

According to Table 1, most of the studies applied geometric perturbations to the fundus images before model training occurred. The common geometric perturbations applied to fundus images are image resize, image crop, image flip, and image rotation. Applying these geometric perturbations to fundus images also benefits from avoiding model overfitting [37].

On the other hand, only three studies adopted photogenic perturbations. Grassmann et al. [17] and Quellec et al. [18] normalized the color balance and local illumination of each fundus image by using Gaussian filtering to subtract the local average color. Gondal et al. [16] transformed the fundus images by manipulating the color, brightness, and contrast of the fundus images. Implementing photogenic perturbations to fundus images can increase the CNN model's performance and correctly detect the lesions present in fundus images. Lin et al. [38] transformed the fundus images by computing spatial entropy to preserve vital characteristics of lesions in the background of fundus images for CNN training.

7 CNN architecture

Various CNN model architectures were adopted in diagnosing DR, as shown in Table 1. Without reinventing the wheel, pre-trained CNN architectures such as AlexNet, DenseNet, and Inception-v4 have been adopted to diagnose DR. Choosing the appropriate CNN model may benefit the interpretability of the CNN model. Quellec et al. [18] reported that the visualization of lesions in fundus images did not perform well in the AlexNet CNN model. They claimed that choosing a CNN model that achieves good image-level performance is preferable to ensure good detection performance at the lesion level. This is true as Gondal et al. [16] adopted an award-winning CNN model architecture o_O network, which performed well in detecting lesions, but had limited capability to detect smaller subtle lesions such as small red dots.

In addition, due to the CNN model architecture where downsampling of features occurred in the convolution layers, information may be a lost, resulting in information loss for very small lesions. Quellec et al. [18] experienced the same phenomena as Gondal et al. [16]. A method proposed by García et al. [39] can be a plausible solution to reduce information loss when downsampling. The authors proposed a heuristic method that identifies zones that are potentially likely to disappear and gives them more importance when filtering. As a result, they are preserved after subsampling.

8 Interpretability techniques

The most commonly adopted interpretability technique in DR-related tasks is the CAM method, as shown in Table 1. The CAM method is popularly adopted due to the use of Global Average Pooling (GAP). The benefit of using GAP is that it acts as a CNN structure regularizer that prevents overfitting during training. Moreover, the GAP explicitly enables a CNN model to have a localization ability despite being trained only on image labels. However, the CAM method has limitations when adopted in DR diagnosis. Although the CAM method was able to highlight areas in the fundus image that contributes to the prediction, the question still arises as to whether the highlighted areas can be trusted. Kumar et al. [21] reported that the CAM method failed to highlight the correct lesions for the misclassified case of mild diabetic retinopathy. Furthermore, Pratt et al. [14] argued that visualization on important lesions areas may not be accurate as there is no pixel-level ground truth used to strengthen the claim. The authors further elaborated that the CNN model is only provided the grades (e.g., classes 0, 1, 2, 3, and 4), not a combination of features and grades, thus deemed unfair that the CNN is expected to precisely highlight the important areas in the fundus image that contribute to the prediction.

In contrast, the occlusion method requires additional model training to observe the diminishing classification score of a particular class. This requirement of re-training a trained model is taxing. To verify the interpretability of the trained CNN model, Kermany et al. [13] and Grassmann et al. [17] systematically blocked different regions of the fundus image with a small gray square box (gray mask) and monitored the output of the classifier, for example, the results in [13] and [17].

Gradient-based interpretability methods such as saliency map, integrated gradient, and layer-wise relevance propagation have their respective limitations as well. The attribution values (heat map) generated using the saliency map method are strongly affected by noisy gradients [25]. While most attribution mass is assigned to the area of the fundus image with the main subject, which seems reasonable, the

Table 3 The summary of qualitative and quantitative metrics used for network's interpretability

Author (year)	Metrics	
	Qualitative visual explanation (heat maps)	Quantitative evaluation
Kermany et al. [13]	✓	✓
Grassmann et al. [17]	✓	✓
Kumar et al. [21]	✓	✗
Tu et al. [15]	✓	✓
Wang and Yang [23]	✓	✗
Gargeya and Leng [20]	✓	✗
Gondal et al. [16]	✓	✓
Jiang et al. [24]	✓	✗
Pratt et al. [14]	✓	✗
Sayres et al. [22]	✓	✗
Quellec et al. [18]	✓	✓
de La Torre et al. [19]	✓	✓

generated heat map is assigned to individual pixels which were affected by high-frequency variations, with neighboring pixels often being assigned very different attributions. Furthermore, a study conducted by Adebayo et al. [40] assessed different gradient-based interpretability techniques to see which visual assessment can be misleading. The strategy appears to rely on the trained model's weights and the link between training instances and their labels, according to the authors.

9 The need for quantitative evaluation

All reviewed literature shows that there are generated heat maps that visually highlight (qualitative) the areas in the fundus image that contribute to the CNN model's prediction. These heat maps act as a medium to allow end-users to understand what discriminative representations the CNN model has learned and how the CNN model makes a prediction. However, we argue that although these generated heat maps can be visually appealing, we still question accurate they are. The evidence shows that many interpretability techniques fail to highlight the correct lesions in a particular class [14–16, 20, 21, 23]. In a classification problem, typically the CNN is trained by only using the fundus images and its respective class labels as input data. There are no explicit lesion features that are fed into the CNN. In addition, the lack of reliable ground-truth medical images is still an open issue in research, thus impeding significant research outcomes [41]. As a result, rigorous quantitative evaluations have not been achieved.

Table 3 summarizes studies that have adopted quantitative evaluation to strengthen the qualitative visual explanations. The CAM interpretability technique lacks quantitative evaluation [14–16, 20, 21, 23]. As the CAM method can localize the discriminative lesion in the fundus image, there should be a quantitative evaluation of the localized area in the fundus image. For instance, scoring unit interpretability called intersection over union (IoU) or Jaccard index metric can be adopted to quantify the localized area in the fundus image [42]. Zhou et al. [43] proposed a method to quantify the interpretability of any given CNN. This proposed method quantifies the interpretability of any given network by measuring the degree of alignment between the unit activation and the ground truth label. This proposed method can be a plausible solution to evaluate the CAM heat maps quantitatively.

10 Conclusion

It is increasingly vital to comprehend and explain sophisticated machine learning models. While various interpretability techniques for DR-related tasks have been adopted, the medical industry's adoption of these techniques continues to be questioned, and research challenges still remain in the scientific community.

To conclude, in this review paper, we examined six interpretability techniques that were adopted for DR-related tasks. The theoretical properties of these techniques have been examined, and it has been demonstrated that, despite their seemingly disparate formulations, they are inextricably linked. In addition, the strengths and limitations of each interpretability technique have been analyzed. The findings have sparked a number of debates in the hopes of spurring more research into new ways for creating explainable deep neural network models.

References

1. WHO. (2020). Diabetes. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes?fbclid=IwAR3prQE7gryQFNPvOhxDorCoIBHqcMFRSOHdHnO3pFN2Gb_V_ipxmQW9MDw
2. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK (2018) Medical image analysis using convolutional neural networks: a review. J Med Syst 42(11):226
3. IDF. (2019). IDF Diabetes Atlas. Available: https://diabetesatlas.org/en/resources/
4. Sommer A et al (2014) Challenges of ophthalmic care in the developing world. JAMA ophthalmology 132(5):640–644
5. Resnikoff S et al (2020) Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? Br J Ophthalmol 104(4):588–592

6. Dean WH, Grant S, McHugh J, Bowes O, Spencer F (2019) Ophthalmology specialist trainee survey in the United Kingdom. Eye 33(6):917–924
7. Mateen M, Wen J, Song S, Huang Z (2019) Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry 11(1):1
8. Gulshan V et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402–2410
9. Arrieta AB et al (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115
10. Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77
11. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5):1–42
12. Pouyanfar S et al (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Computing Surveys (CSUR) 51(5):1–36
13. Kermany DS, et al. (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 5 1122–1131. e9
14. Pratt H, Coenen F, Harding S, Broadbent D, Zheng Y (2019) Feature visualisation of classification of diabetic retinopathy using a convolutional neural network. CEUR Workshop Proceedings 2429:23–29
15. Tu Z, et al (2020) SUNet: a lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) 1378–1382: IEEE
16. Gondal WM, Köhler JM, Grzeszick R, Fink GA, Hirsch M (2017) Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. in 2017 IEEE International Conference on Image Processing (ICIP) 2069–2073: IEEE
17. Grassmann F et al (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
18. Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M (2017) Deep image mining for diabetic retinopathy screening. Med Image Anal 39:178–193
19. de La Torre J, Valls A, Puig D (2019) A deep learning interpretable classifier for diabetic retinopathy disease grading. Neurocomputing
20. Gargeya R, Leng T (2017) Automated identification of diabetic retinopathy using deep learning. Ophthalmology 124(7):962–969
21. Kumar D, Taylor GW, Wong A (2019) Discovery radiomics with CLEAR-DR: interpretable computer aided diagnosis of diabetic retinopathy. IEEE Access 7:25891–25896
22. Sayres R et al (2019) Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126(4):552–564
23. Wang Z, Yang J (2018) Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," in Workshops at the thirty-second AAAI conference on artificial intelligence
24. Jiang H, et al. (2020) A multi-label deep learning model with interpretable grad-CAM for diabetic retinopathy classification. in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 1560–1563: IEEE
25. Ancona M, Ceolini E, Öztireli C, Gross M (2019) "Gradient-based attribution methods," in Explainable AI: Interpreting. Springer, Explaining and Visualizing Deep Learning, pp 169–191
26. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks," in European conference on computer vision 818–833: Springer
27. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034
28. Fong R, Vedaldi A (2019) Explanations for attributing deep neural network predictions," in Explainable ai: Interpreting, explaining and visualizing deep learning: Springer 149–167
29. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition 2921–2929
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision 618–626.
31. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R (2019) "Layer-wise relevance propagation: an overview," in Explainable AI: Interpreting. Springer, Explaining and Visualizing Deep Learning, pp 193–209
32. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365
33. Born J et al (2021) Accelerating detection of lung pathologies with explainable ultrasound image analysis. Appl Sci 11(2):672
34. Jin YW, Jia S, Ashraf AB, Hu P (2020) Integrative data augmentation with U-Net segmentation masks improves detection of lymph node metastases in breast cancer patients. Cancers 12(10):2934
35. Soomro TA, Gao J, Khan T, Hani AFM, Khan MA, Paul M (2017) Computerised approaches for the detection of diabetic retinopathy using retinal fundus images: a survey. Pattern Anal Appl 20(4):927–961
36. Raj A, Tiwari AK, Martini MG (2019) Fundus image quality assessment: survey, challenges, and future scope. IET Image Proc 13(8):1211–1224
37. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. Journal of Big Data 6(1):60
38. Lin G-M, et al (2018) Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy. J Ophthalmol vol. 2018
39. Díaz García J, Brunet Crosa P, Navazo Álvaro I, Vázquez Alcocer PP (2017) Downsampling methods for medical datasets. in Proceedings of the International conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2017 and Big Data Analytics, Data Mining and Computational Intelligence 2017: Lisbon, Portugal, July 21–23, 2017 12–20: IADIS Press
40. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. in Advances in Neural Information Processing Systems 9505–9515
41. Deserno TM, Welter P, Horsch A (2012) Towards a repository for standardized medical image and signal case data annotated with ground truth. J Digit Imaging 25(2):213–226
42. Shanmugamani R (2020). Deep learning for computer vision. Available: https://www.oreilly.com/library/view/deep-learning-for/9781788295628/a5ce2fa2-8c67-4ead-a9bd-a2d07b5f3fa8.xhtml?fbclid=IwAR3pu9MWA93Q1K62qbcJPgpbnPvjKqAljyyprDEUnr8U5D1E9JeGMr0Mwqg
43. Zhou B, Bau D, Oliva A, Torralba A (2019) "Comparing the interpretability of deep networks via network dissection," in Explainable AI: Interpreting. Springer, Explaining and Visualizing Deep Learning, pp 243–252

Wei Xiang Lim , is currently a PhD student at the Department of Science and Engineering, University of Nottingham Malaysia. His PhD research focused on machine learning interpretability in the medical field, specifically in diagnosis of Diabetic Retinopathy.

ZhiYuan Chen , is currently an Assistant Professor with the University of Nottingham, School of Computer Science in Malaysia and a Principal Consultant with MIMOS at the Accelerative Technology Lab. Her research interest is on kernel methods, especially using support vector machines to solve real world problems, such as in anti-money laundering and medical diagnosis.

Amr Ahmed , (BSc'93, MSc'98, PhD'04, MBCS'05), is currently an Associate Professor, School of Computer Science, University of Nottingham, Malaysia Campus. His research focuses on the analysis, understanding, and interpretation of digital contents, especially visual contents. Amr's current research interests include Medical Image/Video analysis, Contents-Based Image/Video retrieval, video and scene understanding, semantic analysis, integration of knowledge and various modalities for scene understanding. Amr worked in the industry for several years, including Sharp Labs of Europe (SLE), Oxford (UK), as a Research Scientist, and other Engineering Consultants companies abroad. He also worked as a Research Fellow, at the University of Surrey, before joining the academic staff at the University of Lincoln in 2005. Dr. Ahmed is a Member of the British Computer Society (MBCS). He completed his Ph.D. degree in Computer Graphics and Animation at the University of Surrey, UK., in 2004