IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 7, 2022, accepted March 1, 2022, date of publication March 8, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157632

Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey

MOHAMMAD Z. ATWANY, ABDULWAHAB H. SAHYOUN, AND MOHAMMAD YAQUB
Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
Corresponding author: Mohammad Z. Atwany (mohammad.atwany@mbzuai.ac.ae)

ABSTRACT Diabetic Retinopathy (DR) is a degenerative disease that impacts the eyes and is a consequence of Diabetes mellitus, where high blood glucose levels induce lesions on the eye retina. Diabetic Retinopathy is regarded as the leading cause of blindness for diabetic patients, especially the working-age population in developing nations. Treatment involves sustaining the patient's current grade of vision since the disease is irreversible. Early detection of Diabetic Retinopathy is crucial in order to sustain the patient's vision effectively. The main issue involved with DR detection is that the manual diagnosis process is very time, money, and effort consuming and involves an ophthalmologist's examination of eye retinal fundus images. The latter also proves to be more difficult, particularly in the early stages of the disease when disease features are less prominent in the images. Machine learning-based medical image analysis has proven competency in assessing retinal fundus images, and the utilization of deep learning algorithms has aided the early diagnosis of Diabetic Retinopathy (DR). This paper reviews and analyzes state-of-the-art deep learning methods in supervised, self-supervised, and Vision Transformer setups, proposing retinal fundus image classification and detection. For instance, referable, non-referable, and proliferative classifications of Diabetic Retinopathy are reviewed and summarized. Moreover, the paper discusses the available retinal fundus datasets for Diabetic Retinopathy that are used for tasks such as detection, classification, and segmentation. The paper also assesses research gaps in the area of DR detection/classification and addresses various challenges that need further study and investigation.

INDEX TERMS Diabetic retinopathy, diabetes mellitus, diabetic macular edema, lesion, microaneurysms, haemorrhages, exudates, classification, supervised learning, self-supervised learning, transformers.

I. INTRODUCTION

Diabetes Mellitus is a chronic disease where blood glucose levels tend to increase due to the lack or inability of the pancreas to produce or secrete sufficient blood insulin [1]. Diabetes incidents have risen rapidly over the past decades, from 108 million in 1980 to 422 million in 2014 [2]. Adverse effects of diabetes on human organs include the liver, heart, kidneys, joints, eyes, etc. [1], [2]. Diabetes serves as the most prominent reason for blindness for people under the age of 50 years. Diabetes Mellitus is a direct cause of Diabetic Retinopathy (DR) which is a complication of diabetes where glucose blocks blood vessels that feed the eye and causes swelling and leaking of blood or fluids that can cause severe eye injury. The detrimental vision loss due to DR occurs primarily when there is retina central swelling. According to the World Report on Vision, an estimated 11.9 million suffer from vision impairment, whether mild or severe, by virtue of glaucoma, trachoma, and DR, which is the focus of our paper [3]. In order to avoid complications associated with chronic diseases such as Diabetes, early detection is vital. Abnormal growth of blood vessels in the retina is a potential consequence of DR, which can cause scarring or bleeding from the retina and consequently blindness [3]. This can result in progressive vision loss with possible blindness at advanced stages. Globally, DR amounts to 2.6% of causes for blindness [4]. The amount of time a patient is diabetic, high haemoglobin A1c, and high blood pressure readings are considered to be the highest risk factors associated with the development of DR [5]. Regular screening is crucial for diabetic patients to ensure that DR is detected at an early stage. DR detection traditionally involves a physician's examination of retinal imaging for the shape and appearance of different types of lesions. Generally, the four types of lesions

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

diagnosed are Microaneurysms (MA), Haemorrhages (HM), soft and hard exudates (EX) [6].

- Microaneurysms (MA) is an early stage of diabetic retinopathy where small red round dots are present on the retina virtue of vessel wall weakness. The dots are defined by sharp margins with a size of not more than 125 micrometres. Microaneurysms can be further classified into six types, but treatment remains uniform regardless of the sub-type [6].
- Haemorrhages (HM) are diagnosed by the presence of large spots on the retina with irregular margin sizes of upwards of 125 micrometres, contrary to Microaneurysms. Haemorrhages can be classified into two categories known as flame and blot, where the spots are superficial and deep, respectively [8].
- Hard exudates are a consequence of plasma leakage and are visible as yellow spots on the retina caused by leakage of plasma. They span the outer retina layers and have sharp margins [1].
- Soft exudates are a consequence of nerve fibre swelling and are visibly white ovals on the retina [6].

Microaneurysms and Haemorrhages commonly appear as red lesions,while the two types of exudates appear as bright lesions. Diabetic Retinopathy detection involves identifying 5 stages which are no DR, mild DR, moderate DR, severe DR and proliferative DR [4], [8]. Figure 1 [7] illustrates the five possible stages of DR development.

The usually occurring retinal lesions incorporate microaneurysms, intraretinal hemorrhages, and venous beading (venous caliber changes consisting of alternating areas of venous dilation and constriction). Furthermore, intraretinal microvascular abnormalities, hard exudates (lipid deposits), and retinal neovascularization [9] are also known to be the common types of lesions.

DR can also be bi-categorized into two main stages known as proliferative DR (PDR) and nonproliferative DR (NPDR). Damage of retina vessels can cause vascular leakage of fluid and circulating proteins into the retina leading to swelling. At this instance, Haemorrhages, Microaneurysms, and exudates can exist for which is known as NPDR. Neovascularization absence determines the diagnosis of NDPR but may tend to include any of the common aforementioned DR lesions. DR progresses sequentially with incremental severity through mild, moderate, and eventually severe PDR that is potentially vision-threatening. Accurate classification of DR severity levels helps identify high risk patients and hence helps mitigating possible problems, through appropriate referrals, cyclic checkups, and proper treatment for sustaining current vision [9].

Proliferative DR represents the latter stages of DR and represents an angiogenic retinal response, in which angiogenesis is a physiological process in which new vessels form from pre-existing blood vessels [10]. Neovascularization of the retina can be commonly viewed as the growth of new vessels along what is referred to as vascular arcades in the retina [9].

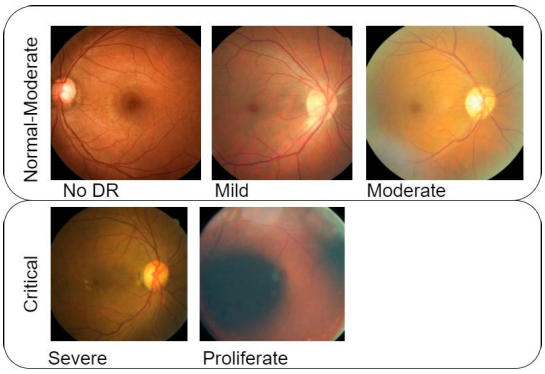

FIGURE 1. The 5 diabetic retinopathy stages, ranked by severity [7].

Manual DR detection requires highly skilled practitioners to perform the assessments. In addition, even highly skilled ophthalmologists suffer from inter- and intra-grader inconsistencies. Therefore, automated detection of DR using accurate machine learning algorithms has the potential to mitigate such shortcomings.

Traditional screening of retinal diseases requires multiple stages of scans followed by filtration techniques to narrow down the subject samples. Optical coherence tomography (OCT) and spatial domain optical coherence tomography (SD-OCT) are examples of scans performed during the screening stage. The resulting fundus images are then sent for analysis by an ophthalmologist. This process is typically prone to a high level of intra-grader inconsistency, Figure 2 shows how DR grading can vary across experts.

There have been multiple efforts to classify OCT images. For instance, OCT fundus images can be classified using the local binary pattern (LBP) proposed in 1990 and enhanced in 2015 by Silva *et al.* [11], but such images are not sufficient for distinguishing between proliferative and non-proliferative DR cases. In retinal imaging, multi-color laser and infrared are used to enhance OCT outputs, as a result the fundus images can be classified with much higher accuracy. This technique allows the detection of lower level abnormalities such as optic discs, but is still not enough for proper DR classification. Emphasis has also been put towards effective image processing techniques as proposed by Gharaibeh *et al.* [12] to further enhance model performance. In computer aided diagnosis (CAD), features of exudates and hemorrhages are highly detectable. This allows fundus images to be clustered into proliferative and non-proliferative cases, where mild and severe vessel abnormalities are distinguished from low level less critical lesions. CAD is one of the fundamental diagnosis techniques in the medical industry [13] and has paved the way for digital medicine. Figure 3 illustrates the different DR lesions that are detectable.

In this paper, we survey the most recent papers related to DR classification. Overall, the focus of this paper highlights the prevalence of Deep learning techniques for DR classification and its impact on classification results.

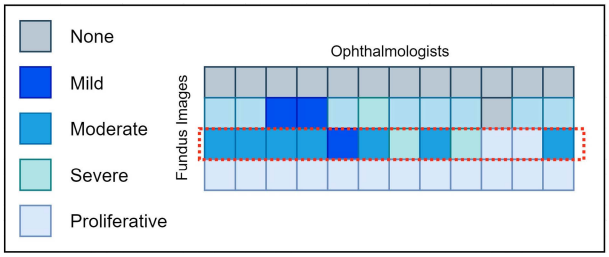

FIGURE 2. Inter-grader inconsistencies illustrated. In the highlighted red area, columns represent a single fundus image and the rows represent the final grade provided by the ophthalmologist.

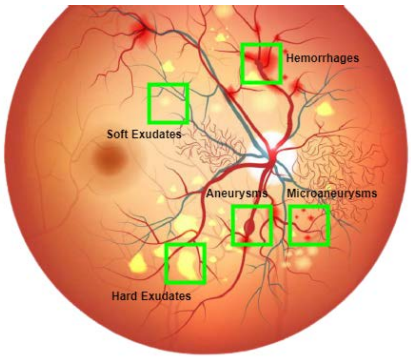

FIGURE 3. A fundoscopic illustration of the retina, showing Microaneurysms, Hemorrhages, and Exudates.

II. REVIEW OF SURVEY PAPERS IN LITERATURE

Attia *et al.* [14] survey examined DR classification methods with a general focus on deep learning techniques and a high focus on classical methods. Gupta and Chhikara [15] reviewed DR detection techniques utilising Adaboost, Random forest, SVM etc, gradually showcasing the gap that these classical techniques present in regards to learning more disease related features. These comparisons are based on quality of the fundus image, since some publicly available datasets have poor contrast and image quality. Alyoubi *et al.* [16] reviewed a total of 33 papers that use deep learning for DR classification and reiterate the importance of constant improvements to deep learning models given the increase in diabetes cases worldwide. Authors also highlighted the use of data augmentation to reduce overfitting in model training. Stolte and Fang [17], Attia *et al.* [14] and Asiri *et al.* [18] survey papers later reviewed novel DL pipelines and ML processes, discussing different DR grading tasks (i.e. optic disc, blood vessels, lesions, and grading). Valarmathi and Vijayabhanu [19] discussed recent state-of the-art(SoTA) CNN variants for DR classification while highlighting the inconsistency in evaluation metrics for the assessment of models in literature. Shamshad *et al.* [20] provides a comprehensive overview of how transformers work for various medical imaging objectives, including: segmentation, classification, detection, and reconstruction. The survey highlights that transformer-based research for medical imaging reached its peak around Dec 2021, with more than 40 recent publications. The survey also shows that 73% of the papers published in 2021 use vision transformers for segmentation tasks whereas 27% of the papers published between 2012 and 2015 use CNNs. This indicates higher demand for transformer-based approaches in segmentation tasks. In terms of retinal diseases, [20] surveys at a few ViT works that target DR grading and classification with lesion detection.

Our review paper sheds light on the DR classification using deep learning techniques. We review papers that also address self-supervision and transformer-based approaches which look to reduce the reliance on large annotated data. Our main methodology review section branches into supervised, self-supervised, and transformer techniques in literature.

III. DATASETS

Retinal fundus images (RFI) are obtained from publicly available standard sets such as DRIVE [21], EyePACS [22], APTOS [23], STARE [24], DIARETDB [25], HEIMED [26], ROC [27], Messidor [28], e-ophtha [29], DDR [30], and RFMiD [31]. The 9 sets are used for comparing different DR classification techniques. In more focused studies, private datasets are leveraged to enhance the accuracy of pre-trained models. Private sets are usually small and typically obtained from participating labs that collaborate with the researchers. Such datasets are not shared as they are kept private. Table 1 provides an overview of all DR open source datasets. Most datasets utilized for training come from EyePACS [22] and Messidor-1 & 2 [28]. All lenses used to capture the fundus images are wide lens CANON cameras with $45° - 50°$ field of view (FoV). The largest datasets used are EyePACS, with 88,702 images [22] and DDR [30], with 13,673 images.

IV. REVIEW OF METHODOLOGY

Diabetic Retinopathy classification can be categorized to either binary classification which aims to detect the presence or absence of DR and multi-class classification, which determines the exact stage of DR. Consequently, further methods were developed to focus on lesion-based classification. Those classification tasks are reviewed under supervised and self-supervised learning in the coming parts of the paper.

A. SUPERVISED METHODS

1) BINARY CLASSIFICATION

Xu *et al.* [45] proposed a DL model to explore the use of CNNs for classification of retinal fundus images with stochastic gradient descent as an optimizer. Authors have experimented with different (9 to 18) layers with varying kernel sizes from (1 to 5). Fundus images were resized to (224,224,3). The model was trained on the Kaggle EyePACS [22] image dataset for normal and Diabetic Retinopathy images. Image rescaling, rotation, flipping, shearing, and translation were used as augmentations to increase the diversity of images needed to train the model and reduce overfitting. In total, 800 training images and around 200 testing images were used. The optimal architecture was based on eight 2D convolutional layers, with max-pooling layer (total

TABLE 1. Datasets used for training DR detection & classification models. Label count represents {N,DR,MDR,SDR,PDR}.

Paper Ref.	Dataset.	Img. Count	Img. Size(px)	Label count	Train+Val/Test Size	DR Grading	Camera Used	No. of studies
[32] [33] [34]	EyePACS 2015 [22]	88,702	433 × 289 to 5184 × 3456	{25810, 2443, 5292, 873, 708}	{35126,53576}	Yes	4 Cameras	3
[33]	APTOS 2019 [23]	3,660	Varies	-	-	Yes	Varies	2
[35] [33] [34]	Messidor [28]	1,200	1440x960 to 2304x1536 (24-bit)	-	-	Yes	45° wide view	3
[35] [36] [37]	Messidor-2 [28]	1,748	1440x960 to 2304x1536(24-bit)	-	-	Yes	45° wide view	3
[35]	IDRiD [38]	516	4288x2848	-	{413,103}	Yes	50° wide view Kowa VX-10	1
[39]	DRIVE [21]	40	565x584 (24-bit)	{33,7}	{20,20}	No	45° wide view CANON CR5	2
[37] [40]	DIARETDB1 [41]	89	1500x1152 (24-bit)	{5,84}	{28,61}	Yes	50° wide view	2
[37]	DIARETDB0 [25]	130	1500x1152 (24-bit)	{20,110}	Varies	Yes	50° wide view	1
[42]	ODIR [43]	10,000	-	{1620,8380}	{9000,1000}	Yes	45° wide view (42 cameras)	1
[42] [37]	DDR [30]	13,673	Varies	{6266, 6256}	{9568,4105}	Yes	Topcon	2
[44]	RFMiD [31]	3200	2144x1424	-	{1920,1280}	No	3 Cameras	1

of 4 max-pooling layers) after 2 convolutional layers, connected at the end to two fully connected layers with a softmax activation function for classification. The model was based on extracted features, namely hard exudates, red lesions, micro-aneurysms and blood vessel detection. To highlight the main contribution of the paper, the Gradient Boosting trees based method coupled with the extracted features mentioned above (Hard exudates + GBM, Red lesions + GBM, Micro-aneurysms + GBM and Blood vessel detection + GBM) was compared against the CNN based model with and without augmentations. GBM hyperparameters were used for the number of classes which was set to 2 with a maximum depth of 6. Extreme Gradient Boosting method (XGBoost) was used due to its superiority against other classical approaches in the literature. ''MXNet'' framework in the R programming language were used.

Quellec *et al.* [40] trained three CNNs to classify retinal fundus images as referable Diabetic Retinopathy for stages 2,3, and 4 and non-referrable Diabetic Retinopathy for stages 0 and 1. Kaggle DIARETDB1 [41] and private E-ophtha images were used for training and evaluation. Images were resized and then (448, 448) crops were taken in the pre-processing stage followed by pixel normalization and then the Field of View (FOV) was eroded by 5% with a gaussian filter applied. The model's architecture was a pre-trained version of AlexNet and two networks of Team o_O solutions for the Kaggle challenge [22]. Microaneurysms, hemorrhages, soft and hard exclusions, and no DR were the classes classified by this model.

Jiang *et al.* [46] used the Adaboost algorithm for efficient integration of several deep learning model outputs using learned weights. Additionally, class activation maps were generated using the outcome of the AdaBoost algorithm and learnt weights in the same manner. The model utilized pre-trained CNNs using Inception V3 [28], Inception-Resnet-V2 [27] and Resnet152 [10] for private dataset classification as referable or non-referable. All their CNN's used Adam optimizers and the AdaBoost algorithm was used to integrate the output of the three CNNs. For data pre-processing, the dataset images were resized to (520, 520, 3) pixels followed by enhancement and augmentation before being used for training. The model was trained using demographically specific data from Chinese population obtained through the Beijing Tongren Eye Center with a total of 30244 fundus images (12513 male cases and 17731 female cases), for diabetic patients between 8 to 98 years of age. For pre-processing, images were normalized to 520 by 520 after cropping the imaging area. After that, an original and transformed fundus image are filtered and combined through weighted summation, providing an enhanced image for different lighting conditions. The transformations applied before training involved translation, rotation, mirroring, brightness, contrast and sharpness. An Adam optimizer was used for the optimization of all three models with a fixed learning rate of 0.001, an exponential decay rate of first-order moment estimation, and second order moment estimation of 0.99 and 0.999, respectively. The model uses local minimums generated by the three sub-models to find the global minima

using the Adaboost algorithm, by decreasing the bias of each respective classifier. The implementation follows three main stages, namely distribution, initialization followed by iterative learning and model combination.

Zago *et al.* [37] utilized augmented image patches of size 65 by 65 pixels for red lesion Diabetic Retinopathy using a pre-trained VGG16 [47] and CNN with five 2D Convolutional layers, five max-pooling layers and finally a Fully Connected layer. Training dataset was DIARETDB1 [41] and the testing was done on DDR [30], IDRiD [38], Messidor-2, Messidor [28], Kaggle [22], and DIARETDB0 [25] datasets for classification of red and non-red lesions. Lesion probability map of the test cases was used to classify images as diabetic or no DR.

2) MULTI-CLASS CLASSIFICATION

This section reviews the studies in which the DR dataset was classified by severity level *i.e. Normal, Mild, Moderate, Severe, & Proliferate*.

Abràmoff *et al.* [36] introduced a method to Diabetic Retinopathy by a CNN model. In pre-processing, the images were normalized and then had a 299 pixel width for the diameter before feeding the images into their model. The model involved training 10 CNNs based on a pre-trained Inception-v3 [48] architecture. The classification involved 5 classes, namely, referable diabetic macular edema, moderate or worse DR, severe or worse DR, or fully gradable. The system put forward by Zhang *et al.* [49] was used to detect Diabetic Retinopathy on their dataset. Their dataset was divided into four classes with a total of 13,767 images. Cropping, resizing, histogram equalization and adaptive histogram equalization were used to pre-process the images. Image enlargement was done through augmentation followed by contrast improvement by a contrast stretching algorithm that is used for dark images. Pre-trained CNN architectures ResNet50 [50], InceptionV3 [48], InceptionResNetV2 [51], Xception [52], and DenseNets [53] were finetuned for Diabetic Retinopathy classification. New fully connected layers were trained on top of the aforementioned CNNs. Then, the pre-trained CNN layers were fine-tuned for retraining, followed by strong model integration. A model utilizing R-FCN with modifications was proposed by Wang *et al.* [54] for detecting stages of Diabetic Retinopathy for the Messidor dataset [28] and their private dataset as well. Their modifications involved modifying the R-FCN by the addition of five region proposal networks and a feature pyramid network. Augmentation was done on training images, with excessive augmentation, particularly for the private dataset images.

Referable and non-referable classification was done on images by Li *et al.* [35] for the Messidor dataset [28]. While the public IDRiD dataset [38] images were classified into five classes (class 0 to class 4) by using four attention modules and ResNet50 [55]. The features extracted by ResNet50 were used as inputs for the initial attention modules. Average pooling, max-pooling, multiplication, concatenation, 2D convolution and fully connected layers are all present in the initial two attention modules. However, the latter two attention modules only contain multiplication and fully connected layers. Image pre-processing included augmentation, normalization, and resizing.

Pao *et al.* [32] utilized bi-channel neural networks for the extraction of fundus components by channel, followed by detail enhancement using a classical sharpness enhancement tool named unsharp masking (UM). The Kaggle Diabetic Retinopathy dataset [22] was used for this implementation, with 21,123 RGB fundus image sizes being selected for this implementation. The images were resized to (100, 100, 3). Then, flipping and rotation are used to yield a total of 33,000 images for the experiments. To compose the 30,000 fundus images for the training set, 15,000 samples are randomly chosen from the first group of grade 0 and another 15,000 from the second group containing fundus images of grade 1 to 4. In the same manner, 3,000 images are chosen for the test set. The bi-channel CNN used for feature learning of referable Diabetic Retinopathy is trained by utilizing features fundus's green component and gray level's entropy images that are initially pre-processed by Unsharp Masking (UM) for enhanced detection of Diabetic Retinopathy especially the referable type. The Unsharp Marking technique is used for the amplification of gray-level high-frequency parts and the green component of the retinal image. Per channel, four convolutional layers are used with 5 by 5 kernels with feature map sizes/number of filters of 32, 64, and 128 for each successive convolutional layer, respectively. For each layer, maximum pooling, rectified linear unit activation function (ReLU) and dropout layers are used, with dropout set to 0.3 This is followed by flattening for the two channels, and the fully connected layers linkage is used to determine the classification of referable DR statistically.

Tymchenko *et al.* [33] used a multi-task learning approach to classify DR classes, by using a deep CNN architecture with a small decoder, namely, head and a feature extractor. Kaggle EyePACs dataset [22] was used for pre-training of the CNN. Other datasets that were combined for the training set were the IDRiD dataset [38] containing 413 photographs of the fundus and the MESSIDOR dataset [28] which contains 1200 fundus images. Augmentations that were performed on the images include optical distortion, grid distortion, piecewise affine transform, horizontal flip, vertical flip, random rotation, random shift, random scale, a shift of RGB values, random brightness and contrast, additive Gaussian noise, blur, sharpening, embossing, random gamma, and cutout. The model uses ImageNet pre-trained CNNs for the initialization of the encoder. They use three decoders in which each decoder is trained to solve its own task virtue of the extracted features using the CNN backbone with classification, regression, and ordinal regression heads. Whereas the classification head output is a one-hot encoded vector where a value of 1 represents the existence of each respective stage. The Regression head's output is a real number in the range from 0 to 4.5 rounded to indicate the different disease stages. As for the ordinal regression head, data points in a category

are all inferred to fall in all categories, thus predicting all categories until the target category. Using an ensemble of three heads and fitting a linear regression model to the outputs of three heads yields the overall prediction. This ensemble is based on the sequential nature of the disease which has been evaluated on Kaggle APTOS 2019 dataset [23].

B. SELF-SUPERVISED METHODS

Supervised learning methods are not the best choice for every problem, especially when the data is noisy. SSL methods are a great alternative to supervised methods and can be used to complement supervised methods. SSL methods are less prone to inductive bias and can be used to handle cross-domain inputs. The problem with SSL methods is that they require a lot of data to be effective. This is a problem because the more data you have, the more time you need to train the model.

1) BINARY CLASSIFICATION

Luo *et al.* [56] introduced a Self-Supervised Fuzzy Clustering Network (SFCN) that is represented by three main modules: a feature learning module for unlabelled retinal fundus images, a fuzzy clustering module for self-supervision and a reconstruction module. Initially, convolutional layers for feature representation extraction make up the feature learning module given an input fundus image, followed by deconvolutional layers for the reconstruction of the retinal images. Adequate information needed to reconstruct the retinal images is satisfied in this module. Then, the feature learning module is provided with training supervision by the fuzzy self-supervision module through the predictions of the fuzzy clustering utilized module algorithm. Through using these self-supervised models, the correlation in unlabeled retinal images is inferred via fuzzy clustering, with probabilities belonging to each respective cluster.

The feature learning module was devised using ResNet50 [55] having one convolutional layer, four residual blocks and ending with a fully connected layer. The Cyclic GAN [57] vanilla image decoder was used to achieve the image decoder, in which the model architecture proposed by Johnson and Fei-Fei [58] constituting a residual block with two stride-1/2 deconvolutions for upsampling purposes with an attached instance normalization. Regarding the fuzzy self-supervision module, two fully connected layers are used following the feature learning module to output the predictions, the Fuzzy C-means clustering output is completely included in the feature module extracted from the convolution layer stack.. During training and testing, images are resized to (224, 224, 3) with a batch size of 32. They used SGD optimizer with an initial learning rate of 0.001 decayed to 0 by the end of the 300 training epochs.

2) MULTI-CLASS CLASSIFICATION

He *et al.* [34] introduced a novel Category Attention Block (CAB) to experiment with features based on regions for each respective DR grade. This network is commonly used for DR multi-class classification in order to mitigate the DR grade imbalance in distribution within most publicly available datasets like Messidor, EyePACS, and DDR. Category attention is used to complement spatial and channel attentions to allow the CAB to be embedded with varying non-category centric blocks for the improvement of multi-class classification, specifically DR grading for this application. The model combines the aforementioned Cabinet with GABNet which is inspired by Woo *et al.* [59] i.e. GAB and CAB, where CABNet is proposed for DR grading. GAB can learn global class-eccentric features while ignoring features like color and texture. In conjunction, CABNet captures detailed features of small lesions to tackle the problem of imbalanced data distribution. Four parts build the CABNet module, namely the backbone, Global Attention Block (GAB), Category Attention Block (CAB) and a classifier. The attention module consists of the GAB and CAB for which the CABNet training is followed end-to-end.

Input fundus images are fed to the CABNet for which the backbone network is merely used to obtain and extract feature maps on a global scale. The model is flexible and thus any CNN architecture can be used for the backbone for which the features can be extracted from the last convolutional layer with highly rich semantic features of the input fundus images. The feature map obtained from the backbone in the earlier step is then initially fed into a 1 by 1 convolutional layer for input channel reduction which is then given as input to GAB and the output of spatial attention is fed as input to CAB and finally to a classifier for DR grading. For training, the base CABNet model has a backbone network pre-trained on the ImageNet dataset. Data transformations applied include random horizontal flips, vertical flips, and random rotation with input images of size (512, 512, 3). The learning rate was initially set at 0.005 and systematically decayed using a factor of 0.8 based on validation loss. Training is performed for 70 epochs using an Adam optimizer and cross-entropy loss function. Different backbone models were trained and the best performing model with the minimum validation loss is used as the base model. The batch size was set to 16.

Lin *et al.* [42] introduced a module named MCG-Net that is based on Graph Convolutional Network (GCN) for the efficient feature extraction of fundus image lesions used for multiclass classification and improved the classification of lesions. To improve generalization, an enhancement module of the introduced MCGS-Net is constructed based on Self Supervised Learning (SSL) in which a GCN is used in place of a fully connected layer to better capture the correlation of fundus images as a classifier. The use of self-supervised learning leads to improvements in the CNN generalization ability. The model has three main components which are the backbone module for sharing feature extraction, CGCN module (GCN for Classification), and GSSL module (SSL for Generalization). Self-Supervision and ODIR datasets are fed to the CNN for image representation extraction. After the global max-pooling layer, a feature vector is obtained. Subsequently, the GSSL allows the MCGS-Net to train on more unannotated data by utilizing the self-supervised

technique for which the GSSL uses a fully connected layer as a classifier. The CGCN then uses the classifier from the GCN for obtaining the category correlation between fundus images. The model uses a total of three datasets for the pipeline, namely the ODIR dataset [43], SSL dataset, and GTest dataset. For testing and training the multi-class network, the open-source ODIR dataset was used. The SSL dataset was used to train the MCGS-Net in a self-supervised manner in which the human-annotated labels were removed. The GTest dataset was then used for testing the generalization ability on six different networks in ablation, mainly MCGS-Net and MCG-Net. The two networks demonstrated better performance over ResNet50 [55], DenseNet121 [60], EfficientNet-B0 [61].

C. TRANSFORMER METHODS

The vision transformer (ViT) proposed by [62] was the first to re-purpose the attention mechanism used for text based data on images. In ViT, image inputs are split into patches of (16,16), projected to embeddings with positional markers, and are then sent to the transformer encoder layers. In the encoder, the process works similar to text-based data, the patch embeddings pass through a multi-head self-attention block, where local and global dependencies of the image are learned and concatenated. The result of the self-attention layer is then normalized for generalizability and sent to the final MLP head, where classification occurs and many papers have utilized their function for Diabetic Retinopathy classification.

Sun *et al.* [63] contributed to bridging the gap between DR grading and lesion discovery by introducing a novel lesion-aware transformer (LAT) using a unified model using a pixel relation based encoder and a lesion filter-based decoder in a weakly supervised lesion discovery localization setup that uses image-level labels only. The model's ResNet50 [55] backbone is mainly for feature extraction for which the fully connected layer and global average pooling (GAP) layers are removed. Then, images are resized to a size of (512,512) and then augmentations such as vertical flips, horizontals flips, random cropping, and color jitter are used to increase the number of images for training to decrease overfitting.The testing is then done on Messidor-1 [28], Messidor-2 [28] and EyePACS [22] datasets.

Kamran *et al.* [64] proposed a SoTA novel conditional generative adversarial network (GAN) that synethsizes Fluorescein Angiography (FA) from fundus images, for which the former is an exogenous dye used to injected in the bloodstream to image the retinal vascular structure and showcase retinal degeneration. The model offers a non-invasive alternative approach while obtaining the benefits of the dye by using a semi-supervised training GAN setup, using different losses with respective weights. This ViT based generative adversarial network (GAN) constitutes residual, spatial feature fusion, upsampling and downsampling modules for the generator, while making use of transformer encoder blocks for the discriminators. The model uses normal and abnormal fundus training images to generate fluorescein angiography (FA) images. For training, the original images of size (576,720) are used to extract 50 images from each respective set with a crop size overlapping of (512,512). As for image synthesise training, a total of 850 images are extracted. Fundus images have 3 channels (RGB) while the FA images have just a single channel. The training and testing datasets are private and split into abnormal and normal classes for which the annotation is used for the supervised classification training part. In total, there are 17 images in the private dataset, of which ten are abnormal and seven are normal patients. As aforementioned, due to cropping, this then extends to 500 images for abnormal class and 350 for the normal class.

Papadopoulos *et al.* [65] proposed a transformer-based method for the independent extraction of required local information through combining several rectangular patches with an efficient attention structure focused on eye regions with lesions (abnormal images) for classifying images. The model also utilizes the attention mechanism to generate heatmaps. The image preprocessing process included Hough transform, random resized cropping, subtracting image local color average for different lighting conditions, and zeroing the outer 5 percent of the retina disk. The model was trained on the kaggle EyePACs datset [22] andd testing conducted for Messidor-2 [28], IDRiD [38] and Kaggle EyePACs [22] datasets as well.

Yu *et al.* [44] introduced a novel transformer-based technique through pre-training on a large number of fundus images followed by fine-tuning on a classification task. The model uses a multiple instance learning (MIL) based 'MIL head' that is attached to the ViT in a plug-and-play manner to improve on the downstream classification task.The private dataset used for training mainly was obtained from a tele-opthalmology platform, which contains a total of 345,271 fundus images with the most common retinal conditions labeled namely normal, diabetic retinopathy, glaucoma, cataract and macular degeneration, but since some conditions occur simultaneously the pretraining is set as a multiclass classification task with 95 and 5 percent split for training and validation, respectively and images resized to (384,384). As for the downstream task the APTOS [23] and RFMiD [31] datasets are used for training and testing. For the two datasets, the images are resized to (512,512).

V. REVIEW OF RESULTS

This section sheds light on the most novel studies with their reported results. The idea is to compare and discuss the top performing methods and explore their scalability and generalizability to some extent. All results are reported in Table 3.

A. EVALUATION METRICS

For DR detection and classification, the AUC, F1, and Kappa scores are used as the main metric for determining the validity of the results, while other metrics such as accuracy and recall can be considered support metrics. This is because data distribution is imbalanced within each dataset. The equations

TABLE 2. Evaluation metrics.

Metric	Formula.	Methods	Usage
Accuracy	$\frac{TN+TP}{TN+TP+FN+FP}$	Binary(Normal/Abnormal)	Where mis-classification is not critical.
Recall	$\frac{TP}{TP+FN}$	Binary(Normal/Abnormal)	Cost of False Negative is extremely high. May lead to critical DR cases.
Specificity	$\frac{TN}{TN+FP}$	Binary(Normal/Abnormal)	Inversely related to Recall. Highly specific test that can determine Abnormal DR cases (True Negative is important).
Precision	$\frac{TP}{TP+FP}$	Multi classification	Cost of False Positive is high. The DR stage must be determined with precision.
F1	$\frac{2*Prec*SN}{Prec+SN}$	Multi classification	Uneven DR class distributions and provides a balance between precision and recall.
Kappa	$\frac{P_0-P_e}{1-P_e}$	Multi classification	Uneven DR class distribution and accounts for ground truth labels across all classes.

in Table 2 show how the metrics are calculated with some use-cases.

B. SUPERVISED RESULTS

1) BINARY CLASSIFICATION

Pao *et al.* [32], presents a model that generates entropies of each fundus image, allowing it to further highlight the lesion edges and create areas of interest for the feature extractor during binary classification. With that, it achieves 87.37% accuracy and an F1 of around 81.8%. In this study, the AUC of the ROC curve is used as a core indicator of performance, ignoring imbalances in the fundus data. Both studies by Tymchenko *et al.* [33] and Pao *et al.* [32] extract features from lesions to some extent, but fail to identify the type of lesion with the final output. Tymchenko *et al.* [33] takes it one step further by identifying the DR stage as being one of the five stages: mild, moderate, severe, or proliferate.

Similarly, the model proposed by Zhang *et al.* [49] was able to identify the different severity levels and achieve a multi-class accuracy, specificity and sensitivity of 96.5%, 98.9% and 98.1%, respectively. However, this model does not detect retinal fundus lesions and the results were limited to the private dataset used for evaluation.

Hua *et al.* [39] proposed an uncommon model RFA-BNET that stands out due to its approach, however it utilizes a ResNet-101 as it's backbone. Hua *et al.* [39] model aggregates features from multiple rounds through the ResNet-101, it achieves a relatively lower accuracy rate than the rest of the ensemble models. Hua *et al.* [39] reported a 95.1% accuracy and a recall of 79.3%.

2) MULTI-CLASS CLASSIFICATION

The study by Tymchenko *et al.* [33], utilizes transfer learning in a 3-headed ensemble CNN architecture (classification, ordinal, and regression) and achieved the best results using an ensemble of 20 different models and a trimmed mean of 200 five-class predictions for each fundus image. Using training time augmentations (TTA), the model achieved 99.3% accuracy. The model's quality was assessed using binary classification screening and achieved an F1 score of 99.3%.

One of the most complex models proposed by Zhang *et al.* [49] utilizes an ensemble of pre-trained networks that is ultimately developed into a whole framework for DR detection, but when compared to Li *et al.* [35] study which uses a single ResNet50 layer, the results are not that far from what can be achieved by a high-grade ensemble network. Li *et al.* [35] was able to achieve a sensitivity, AUC, and accuracy of 92%, 96.3% and 92.6%, respectively, for the Messidor dataset [28], while the achieved accuracy for IDRiD was around 65.1%, no precision or F1 was reported. In an attempt to reduce the number of trainable parameters, Zago *et al.* [37] proposes a a VGG16 network achieving a sensitivity of 0.94 and an AUC of 0.912 for the Messidor dataset [28], in contrast, this model can only detect DR without any indication of severity. Similarly, no precision or F1 was reported.

C. SELF-SUPERVISED RESULTS

1) BINARY CLASSIFICATION

Luo *et al.* [56] uses the SFCN model trained on 25 DRIVE images. It was able to achieve an accuracy of 81.7%. In comparison, Hua *et al.* [39] RFA-BNET uses 20 images with extensive augmentations, and achieves an accuracy of 95.1%. No tests were reported to show its generalizability to other sets like EYEPACs.

2) MULTI-CLASS CLASSIFICATION

CABNet by He *et al.* [34] trained with 70 epochs on the EYEPACS set, was able to generalize to Messidor and achieve a very competitive accuracy and kappa of 84% and 85.5% respectively. CABNet's custom-made attention block makes it easier to interpret how the model works and what areas of the fundus images contribute to the embeddings.

TABLE 3. DL methods for DR detection & classification. Highest scores reported.

Paper Ref.	Lesion Det.	Classifier	Arch.	SSL	Data set	Acc	Recall (SN)	SP	AUC	F1	Kappa
Supervised CNN											
[45]	No	Binary	CNN	No	EyePACS	94.5%	-	-	-	-	-
[40]	No	Binary	CNN AlexNet	No	EyePACS, DIARETDB1	-	-	-	95.4% and 94.9%	-	-
[46]	No	Binary	Inception-V3 & ResNet152	No	private dataset	88.21%	-	-	94.6%	-	-
[37]	Yes	Binary	VGG16	No	DIARETDB1	-	94%	-	91.2%	-	-
[36]	No	Multi	Inception-V3	No	EyePACS	-	97.5%	93%	-	-	-
[49]	No	Multi	Multi-CNN	No	private dataset (13,767)	96.5%	98.1%	98.9%	98.62% (Ensemble)	95.42%	-
[54]	Yes	Multi	CNN	No	private dataset (9,194), Messidor (1,200)	92.95%	99.39%	99.93%	-	-	-
[35]	Yes	Multi	Resnet50	No	Messidor, IDRiD	92.6%	92%	-	96.3%	91.2%	-
[32]	Yes	Binary	Bi-channel CNN	No	EyePACS	87.37%	76.93%	93.57%	93%	-	-
[39]	Yes	Binary	CNN (ResNet-101)	No	DRIVE	95.1%	79.3%	97.4%	97.32%	-	-
[33]	No	Binary + Multi	3-Headed CNN (TTA)	No	private dataset (736), EyePACS (13,000), APTOS	91.9% (Multi)	84.0% (Multi)	98.1% (Multi)	-	84.0% (Multi)	-
Self-Supervised CNN											
[56]	Yes	Binary	CNN + Fuzzy C-Means	Yes	DRIVE, Messidor, DIARETDB1	95.1%	-	-	-	-	
[34]	Yes	Multi	CNN + CAB & GAB	Yes	EyePACS	84.0%	-	-	-	-	85.5%
[42]	Yes	Multi	CNN + GSSL & GCN	Yes	ODIR, SSL, GTest	-	-	-	-	86.56%	38.77%
Transformers											
[63]	Yes	Multi	CNN+ ViT	No	Messidor-1&2, EyePACS	-	-	-	98.7%	-	88.4%
[64]	No	Binary	GAN+ ViT	No	private dataset	85.7%	83.3%	90.0%	-	-	-
[65]	Yes	Binary (rDR)	ResNet-18+ Attention (MIL+Ensemble)	No	Messidor-2, EyePACS	-	99%	92.0%	99.0%	-	-
[44]	No	Binary + Multi	ViT (MIL)	No	APTOS, RFMiD	85.5% (Multi)	93.7% (Binary)	-	97.9% (Multi)	85.3% (Multi)	92.0%(Multi)

Such an architecture gives CABNet an edge in terms of customizing the attention blocks and adds to its flexibility.

Lin *et al.* [42] MCGG-Net on the other hand, is trained with 60 epochs on 9000 ODIR images and achieves an F1 and kappa of 86.56% and 38.77% on the GTest dataset respectively. These results show that the model is able to generalize to other fundus images on an entirely different set, using only the embeddings of an image with little to no labels. In comparison, the best performing multi-class supervised model by Tymchenko *et al.* [33], was able to achieve

an F1 and kappa of 84% and 96.9% respectively. This three-headed CNN uses a complex ensemble training process with extensive augmentations. MCG-Net only performs a single augmentation of a flips, while CABNet uses flips and rotations. Given the fact that SSL models see great efficiency in the data pre-processing stage, they are generally faster to fine-tune and adapt to new fundus images and can generalize far better than supervised models. The benefit comes into play where research is required to be done on a wide range of datasets.

D. TRANSFORMER RESULTS

1) BINARY CLASSIFICATION

VTGAN by Kamran *et al.* [64] introduced an GAN evaluation framework that leverages two assessments. The qualitative assessment looks at the architecture performance in terms of Frechet Inception Distance (FID), which looks at the image quality of the GAN model and Kernel Inception Distance (KID) which looks at image features and structural similarity to the original fundus counterpart. The results indicate at least 30% better FID and KID scores compared to SoTA like A2GAN [66] and StarGAN-v2 [67]. In the transformed images, VTGAN also gains a 5% improvement over SoTA. The Overall average qualitative precision score is 45.9% and an average quantitative accuracy of 78% has been achieved on out-of-distribution transformed images vs. 85.7% on in-distribution fundus.

Another Multiple-instance learning (MIL) study by [65], utilizes MIL in its attention mechanism to treat patches of DR lesions as a bag of features, retaining only relevant information for the classifier to work with. This technique allowed the production of high-quality attention maps that highlight detected lesions, eliminating the black box. Using random patch selection, the MIL model [65] achieved 95.7% AUC on Kaggle EYEPACS [22] and 99% AUC on Messidor-2 [28]. The paper also studies the effect of lesions on attention weights using a lesion classification approach. The conclusion was that smaller lesions with more characteristics or variations yielded higher attention weights. AUC of 80% was achieved on all lesions (Microaneurysms, Haemorrhages, and Exudates).

2) MULTI-CLASS CLASSIFICATION

MIL-VT [44] is another similar MIL model that is capable of classifying the DR disease level. The approach used in MIL-VT is almost identical to the vanilla ViT [62] approach for generic images, with the addition of an MIL embedding layer that aggregates patches based on features and attention before sending them to the MIL classifier (MIL Head). On APTOS [23], MIL-VT achieved 94.4% F1 in DR disease classification and 85.5% accuracy in DR grading. Compared to SoTA, GREEN-SE-ResNext50 achieved an F1 score of 85.3% and accuracy of 85.7%.

Lesion Aware transformer (LAT) [63] performs DR grading using attention blocks rather than full transformer architectures. The idea is that a self-attention layer encodes and outputs contextual features of lesions that are then blended into the final output. By using an approach known as lesion importance learning, [63] achieved a DR grading accuracy of 96.3% on normal fundus and 98.7% on fundus with detected lesions (referrals). This outperforms SoTA such as CANet [35] and Semi+Adv [68]. What makes the evaluation unique is the ablation study developed to assess the 5 network components, namely: pixel relation encoder (P), self-attention layer (S) and cross-attention layer (C), region diversity mechanism (D), and global consistency loss (G). By gradually adding the components and testing the network end-to-end using AUC and Kappa metrics, it was concluded that LAT [63] achieved maximum Kappa and AUC using all 5 components working collectively.

VI. DISCUSSION

This paper reviews 11 supervised, 3 self-supervised, and 4 transformer papers by analyzing the results and techniques of each method used. The main idea behind this paper is assess DR grading and classification methodologies from a qualitative point of view, which essentially allow future research work to be aware of the advancements in the DR domain.

In the supervised methods, 54% of the studies reviewed use binary classification for DR detection, while the remaining 46% classify DR by 4 severity stages. In the self-supervised methods, 67% use multi-class detection and the remaining 33% use binary. In terms of datasets, a few studies tried to use self-developed private sets for total control, this generally showed enhanced results as with Tymchenko *et al.* [33] and Zhang *et al.* [49], however, no correlation can be deduced due to varying data distributions. About 57.2% of the studies train their models on multiple datasets while the remaining 42.8% use a single dataset. Figure 4 summarizes some collected statistics. Supervised models are robust and work great for specific tasks. A common issue the papers generally fail to highlight is the extreme amount of time needed to train new models every time. For mission critical models that are constantly bombarded with new data, the process of annotating and structuring the data to make it ''model ready'' almost defies their purpose. Ensemble learning pipelines in DL are great at handling features in variable data distributions and must be promoted in the supervised learning context.

Self supervised models show a competitive edge regarding fine-tuning on new sets. All 3 papers suggested that training on a very large dataset such as Messidor or EYEPACS would allow the model to generalize to a wider range of sets, such as DRIVE and GTest. This was proven to be true from the presented results. However, in order to make sense of SSL models, it is required to understand how it interprets the data that is fed into it. Luo *et al.* [56] Luo's paper uses t-SNE plots to interpret how SFCN is able to partition normal and abnormal images. A. He *et al.* [34] uses attention maps to visualize what the model focuses on in terms of features. The heatmaps illustrate how CABNet's attention block helps narrow down

FIGURE 4. Top chart shows the classification segment distribution across 11 supervised studies (inner circle), 3 self-supervised studies (middle circle), and 4 transformer studies (outer circle). Bottom chart shows evaluation metric usage distribution across 11 supervised studies (inner circle), 3 self-supervised studies (middle circle), and 4 transformer studies (outer circle).

the features that are selected. This can essentially reduce the size of embeddings needed to generalize. While studies show the power of self supervised approaches relative to supervised methods, they do not show how SSL methods can be less prone to inductive bias in the long run. SSL methods are known to cope well with cross domain inputs as well, this sort of advantage is crucial when selecting a model for mission critical applications. Moreover, the papers fail to discuss SSL methods' robustness when working with small-scale datasets.

DR screening remains an open issue due to the limited number of public datasets, and while most of the recent DL advancements achieve promising scores in classification, some still lack the ability to distinguish affected lesions. Other methods just ignore the 5 DR stages that are considered crucial for determining the severity of the disease. This sort of dissimilarity in the techniques reveals another impeding problem. The fact that there are no standard practices that agree on a common set of DR stages shows that researchers in the field may still have different opinions on the validity of the results. Most of the end results, however, can only be used for diagnosis and are never deemed final.

As a future direction, upcoming studies should focus on leveraging SSL methods to not only generalize but also be able to generate new fundus images based on the learned features using generative networks. Generative adversarial networks (GAN) and Variational auto-encoders (VAE) can be combined with existing networks to synthesize a whole range of enhanced fundus images that can be made available for training. As an example, DALL-E proposed by Ramesh *et al.* [69], is capable of generating images from text, such a model could potentially generate large collections of DR fundus images that can be trained and tested on.

Another direction could be the utilization of self-supervised vision transformers, such as DINO proposed by Caron *et al.* [70] to encode better features when large-scale DR sets are provided. Transformers have shown a positive correlation between the number of trainable parameters and accuracy, hence they are immune to saturation with larger sets and varying data distributions. What vision transformers would solve is the increasing complexity of CNN layers as the size of the filters increase. CNNs cannot capture a global understanding of the image because they may not retain visual features throughout the network. Meanwhile, vision transformers take their advantage from their attention mechanisms and are able to find relationships in flattened feature sequences.

In recent studies, attention mechanisms brought upon significant changes to the way images are interpreted and contextualized. The transformer-based models surveyed show promising performance in the medical imaging domain for binary and multi-class classification of DR disease. One notable improvement introduced in these transformer studies, is the ability to distinguish smaller lesions in much more detail, providing better explainability to the classification obtained. While the results show satisfactory performance and improvements over CNNs, there are yet to be additional studies that benchmark models in deployed environments. In most cases, vision models tend to fail once deployed in the real world.

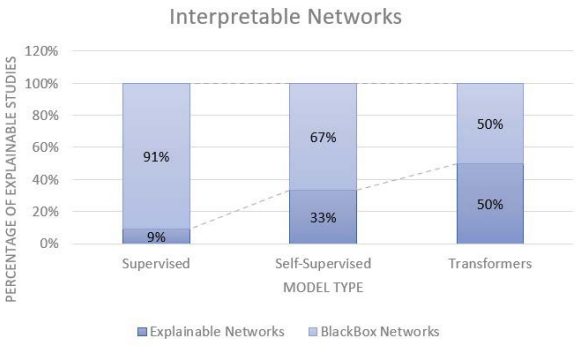

FIGURE 5. Distribution of studies in terms of model explainability, sorted by model type. As transformers evolve, more explainable models emerge.

VTGAN [64] is one example that works really well in theory, but may miss out on certain distortions that are otherwise not typically synthesized in fundus image generation. For instance, noise, blur, and warping may coincidentally exist in real fundus scans, but are not accounted for together in the qualitative assessment, mainly due to architectural limitations in the network.

On the other hand, transformer-based studies introduced more valuable white-box approaches, such as different ablation and evaluation techniques that target explainability and justification of results. By using attention maps and feature enrichment techniques, the models incorporate context more effectively and are able to produce attention maps that highlight detected lesions. Explainable architectures are deemed to be more plausible in the real-world industry. In the medical domain, every decision needs to be justified through insight, research, and scientific proof. Interpretable design is the key in integrating medical imaging models in the end-to-end operational pipelines of many institutions and research facilities.

VII. CONCLUSION

While DR cannot be cured, it is important to detect it in its early stages to prevent further damage. For example, non-proliferative DR stages will almost always contain early indicators of DR and the ability to detect and classify those stages using a proper evaluation technique could mean saving one's eyesight. In this review paper, a major portion of the work focuses on the study of hemorrhages, microaneurysms and exudates. Results from multiple studies show an accuracy average of about 91% and promising classification performance overall. Screening systems being developed today could incorporate these DL based approaches to enhance and classify the DR stage using lesion detection techniques across multiple fundus images. The main issue addressed in the reviewed studies is the manual diagnosis that has to occur after screening, which is typically a lengthy process prone to ophthalmologists' bias. Moreover, dataset limitations restrict fundus image variations that can be used in the assessment of indicators.

Because of the efficiency of Deep Learning techniques, the analysis of retinal scans has become faster, more inclusive, and generalizable, yet the metrics used in the evaluation of the results and their respective datasets remain biased and unbalanced across different studies. Ultimately, classifying DR is crucial, but understanding the various causes can also be a valid research opportunity. For instance, specific lesion changes and other hidden indicators could potentially hint at the possibility of developing DR. Other research directions could involve studying Diabetic Macular Edema (DME) since detecting DME is highly likely to mean that the retina is developing DR. With these advancements, it is possible to generalize DL based models and assess a wider range of symptoms and indicators that could help researchers get a better understanding of the causes of retina based diseases.

Lastly, Transformers introduced more explainable methods that can help overcome the limitations of non-generalizability. Hidden indicators can now be detected more accurately, thanks to the various context enrichment approaches used in patching and embedding of fundus images. In SSL and Supervised methods, only 2 papers stood out in terms of interpretability of the results, namely, Tymchenko's CNN Ensemble model with SHAP analysis [33] and He's CABNet with attention maps [34]. Figure 5 shows the study distribution in terms of method and network interpretability. Ultimately, transformer-based models enable better interpretability for researchers and future work

REFERENCES

[1] J. Amin, M. Sharif, and M. Yasmin, "A review on recent developments for detection of diabetic retinopathy," *Scientifica*, vol. 2016, pp. 1–20, Sep. 2016.
[2] A. T. Kharroubi and H. M. Darwish, "Diabetes mellitus: The epidemic of the century," *World J. Diabetes*, vol. 6, no. 6, pp. 850–867, Jun. 2015.
[3] *World Report on Vision*, World Health Organization, Geneva, Switzerland, 2019.
[4] S. Mamtora, Y. Wong, D. Bell, and T. Sandinha, "Bilateral birdshot retinochoroiditis and retinal astrocytoma," *Case Rep. Ophthalmolog. Med.*, vol. 2017, pp. 1–4, Feb. 2017.
[5] J. W. Yau *et al.*, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.
[6] M. Dubow, A. Pinhas, N. Shah, F. R. Cooper, A. Gan, C. R. Gentile, V. Hendrix, N. Y. Sulai, J. Carroll, Y. P. T. Chui, B. J. Walsh, R. Weitz, A. Dubra, and B. R. Rosen, "Classification of human retinal microaneurysms using adaptive optics scanning light ophthalmoscope fluorescein angiography," *Investigative Ophthalmol. Visual Sci.*, vol. 55, no. 3, pp. 1299–1309, Mar. 2014.
[7] P. Vora and S. Shrestha, "Detecting diabetic retinopathy using embedded computer vision," *Appl. Sci.*, vol. 10, no. 20, p. 7274, Oct. 2020.
[8] N. Murugesan, T. Üstunkaya, and E. Feener, "Thrombosis and hemorrhage in diabetic retinopathy: A perspective from an inflammatory standpoint," *Seminars Thrombosis Hemostasis*, vol. 41, no. 6, pp. 659–664, Aug. 2015.
[9] Y. T. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, V. C. Lansingh, M. Maia, W. Mathenge, S. Moreker, M. K. M. Muqit, S. Resnikoff, J. Verdaguer, P. Zhao, F. Ferris, P. L. Aiello, and R. H. Taylor, "Guidelines on diabetic eye care," *Ophthalmology*, vol. 125, no. 10, pp. 1608–1622, Oct. 2018.
[10] A. Birbrair, T. Zhang, Z.-M. Wang, M. L. Messi, A. Mintz, and O. Delbono, "Pericytes at the intersection between tissue regeneration and pathology: Figure 1," *Clin. Sci.*, vol. 128, no. 2, pp. 81–93, Jan. 2015.
[11] C. Silva, T. Bouwmans, and C. Frélicot, "An eXtended center-symmetric local binary pattern for background modeling and subtraction in videos," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, Mar. 2015, pp. 395–402.
[12] O. M. Al Hazaimeh, K. M. O. Nahar, B. Al Naami, and N. Gharaibeh, "An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images," *Int. J. Signal Imag. Syst. Eng.*, vol. 11, no. 4, p. 206, 2018.

[13] H. Fujita, Y. Uchiyama, T. Nakagawa, D. Fukuoka, Y. Hatanaka, T. Hara, G. Lee, Y. Hayashi, Y. Ikedo, X. Gao, and X. Zhou, "Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs," *Comput. Methods Programs Biomed.*, vol. 92, pp. 238–248, Jun. 2008.

[14] A. Attia, Z. Akhtar, S. Akrouf, and S. Maza, "A survey on machine and deep learning for detection of diabetic RetinopathY," *ICTACT J. Image Video Process.*, vol. 11, no. 2, pp. 2337–2344, Dec. 2020.

[15] A. Gupta and R. Chhikara, "Diabetic retinopathy: Present and past," *Proc. Comput. Sci.*, vol. 132, pp. 1432–1440, Jan. 2018.

[16] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100377.

[17] S. Stolte and R. Fang, "A survey on medical image analysis in diabetic retinopathy," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101742.

[18] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi, "Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey," 2018, *arXiv:1811.01238*.

[19] S. Valarmathi and R. Vijayabhanu, "A survey on diabetic retinopathy disease detection and classification using deep learning techniques," in *Proc. 7th Int. Conf. Bio Signals, Images, Instrum. (ICBSII)*, Mar. 2021, pp. 1–4.

[20] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[21] A. H. Asad, A. T. Azar, N. El-Bendary, and A. E. Hassaanien, "Ant colony based feature selection heuristics for retinal vessel segmentation," 2014, *arXiv:1403.1735*.

[22] *Diabetic Retinopathy Detection EYEPACS Dataset*, Kaggle, San Francisco, CA, USA, Jul. 2015.

[23] *APTOS* 2019 Blindness Detection, APTOS, Atlanta, GA, USA, Jun. 2018.

[24] *The STARE Project*, Shiley Eye Center, San Diego, CA, USA, 2004.

[25] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kalviainen, and J. Pietila, "DIARETDB 0: Evaluation database and methodology for diabetic retinopathy algorithms," Dept. Med., Univ. Kuopio, Kuopio, Finland, Tech. Rep. 573081, 2007.

[26] L. Giancardo, F. Meriaudeau, P. T. Karnowski, Y. Li, S. Garg, W. K. Tobin, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Med. Image Anal.*, vol. 16, no. 1, pp. 216–226, 2012.

[27] M. Niemeijer *et al.*, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[28] E. Decenciere, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[29] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J. C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Lay, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," in *Proc. IRBM*, vol. 34, no. 2, Apr. 2013, pp. 196–203.

[30] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, Oct. 2019.

[31] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau, "Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, Feb. 2021.

[32] S.-I. Pao, H.-Z. Lin, K.-H. Chien, M.-C. Tai, J.-T. Chen, and G.-M. Lin, "Detection of diabetic retinopathy using bichannel convolutional neural network," *J. Ophthalmol.*, vol. 2020, pp. 1–7, Jun. 2020.

[33] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep learning approach to diabetic retinopathy detection," 2020, *arXiv:2003.02261*.

[34] A. He, T. Li, N. Li, K. Wang, and H. Fu, "CABNet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 143–153, Jan. 2021.

[35] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, May 2020.

[36] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Invest. Ophthalmol. Vis. Sci.*, vol. 57, no. 13, pp. 5200–5206, 2016.

[37] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. Teatini Salles, "Diabetic retinopathy detection using red lesion localization and convolutional neural networks," *Comput. Biol. Med.*, vol. 116, Jan. 2020, Art. no. 103537.

[38] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Sep. 2018.

[39] C.-H. Hua, T. Huynh-The, and S. Lee, "Retinal vessel segmentation using round-wise features aggregation on bracket-shaped convolutional neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 36–39.

[40] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, Jul. 2017.

[41] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietila, "The DIARETDB1 diabetic retinopathy database and evaluation protocol," in *Proc. Brit. Mach. Vis. Conf.*, Jan. 2007, pp. 1–10.

[42] J. Lin, Q. Cai, and M. Lin, "Multi-label classification of fundus images with graph convolutional network and self-supervised learning," *IEEE Signal Process. Lett.*, vol. 28, pp. 454–458, 2021.

[43] Larxel. (Apr. 2020). *Ocular Disease Recognition*. [Online]. Available: https://odir2019.grand-challenge.org/

[44] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, *MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification*. Cham, Switzerland: Springer, 2021.

[45] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, p. 2054, Nov. 2017.

[46] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2045–2048.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[49] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowl.-Based Syst.*, vol. 175, pp. 12–25, Jul. 2019.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jul. 2015.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*.

[52] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, *arXiv:1610.02357*.

[53] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.

[54] J. Wang, J. Luo, B. Liu, R. Feng, L. Lu, and H. Zou, "Automated diabetic retinopathy grading and lesion detection based on the modified R-FCN object-detection algorithm," *IET Comput. Vis.*, vol. 14, no. 1, pp. 1–8, 2020.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[56] Y. Luo, J. Pan, S. Fan, Z. Du, and G. Zhang, "Retinal image classification by self-supervised fuzzy clustering network," *IEEE Access*, vol. 8, pp. 92352–92362, 2020.

[57] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*.

[59] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

[60] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.
[61] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[63] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, "Lesion-aware transformers for diabetic retinopathy grading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10933–10942.
[64] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, and S. A. Baker, "VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers," 2021, *arXiv:2104.06757*.
[65] A. Papadopoulos, F. Topouzis, and A. Delopoulos, "An interpretable multiple-instance approach for the detection of referable diabetic retinopathy from fundus images," 2021, *arXiv:2103.01702*.
[66] S. A. Kamran, K. F. Hossain, A. Tavakkoli, and S. L. Zuckerbrod, "Attention2AngioGAN: Synthesizing fluorescein angiography from retinal fundus images using generative adversarial networks," 2020, *arXiv:2007.09191*.
[67] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," 2019, *arXiv:1912.01865*.
[68] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2074–2083.
[69] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021, *arXiv:2102.12092*.
[70] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021, *arXiv:2104.14294*.

MOHAMMAD Z. ATWANY received the B.Sc. degree in mechanical engineering from the University of Sharjah, Sharjah, UAE. He is currently pursuing the M.Sc. degree in machine learning with the Mohamad bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. His applied industrial and commercial experience includes predictive maintenance models for facilities management systems. His research interests include but are not limited to deep learning and machine learning in medical imagining with a focus on generative adversarial networks and variational inference.

ABDULWAHAB H. SAHYOUN was born in Toronto, ON, Canada, in 1994. He received the B.Sc. degree in computer engineering from the American University of Sharjah, UAE, in 2016. He is currently pursuing the M.Sc. degree in machine learning with the Mohammad bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE.

He is the author of one publication at the 2016 IEEE 18th International Conference (Healthcom), on the diagnosis of Parkinson's disease using mobile phones, "ParkNosis." At MBZUAI, he is currently researching Arabic NLP models trained on speech. He has industry and commercial experience in NLP-based data processing systems.

MOHAMMAD YAQUB received the Ph.D. degree from the University of Oxford, Oxford, U.K. He is currently an Assistant Professor with the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. He leads the BioMedIA Group at MBZUAI, where his team investigates the development of machine learning solutions to several real-world healthcare problems. He has teaching experience in higher education alongside years of research experience in world leading institutions. His domain knowledge is artificial intelligence (AI), machine learning, big data, and biomedical image analysis. In addition, he is an University Lecturer with IT Services, University of Oxford, where he is a Research Fellow with the Nuffield Department of Clinical Neurosciences (NDCN). Furthermore, he teaches courses through Oxford "English Medium Instruction (EMI)," to non-English speaking multidisciplinary university lecturers from all over the world. He has co-supervised master's and D.Phil. students at IBME, University of Oxford.