STM JOURNALS

Research & Reviews: A Journal of
Medical Science and Technology

https://journals.stmjournals.com/rrjomst/

ISSN: 2319-3417 (Online)
ISSN: 2349-1272 (Print)
Volume 13, Issue 3, 2024
September–December
DOI (Journal): 10.37591/RRJoMST

Research RRJoMST

The Analysis of Deep Learning-Based Methods for Identifying Diabetic Retinopathy

Tshetiz Dahal[1,*], Suyash Saurabh[2]

Abstract

Diabetic retinopathy (DR) is a degenerative eye condition resulting from diabetes mellitus, where high blood glucose levels lead to lesions on the retina. This condition is considered the leading cause of blindness among working-age diabetic patients, particularly in developing countries. As the disease is irreversible, the treatment aims to preserve the patient's current vision. Early detection is crucial for effective management of DR to maintain vision. One of the main challenges in DR detection is the significant time, cost, and effort required for manual diagnosis, which involves having an ophthalmologist review retinal fundus images. The latter also turns out to be more challenging, especially when the sickness is still in its early stages and the signs of the illness are less noticeable in the pictures. Deep learning algorithms have helped in the early identification of DR, and machine learning-based medical image analysis has demonstrated proficiency in evaluating retinal fundus pictures. To propose retinal fundus picture classification and detection, this study discusses and analyzes the most recent deep learning techniques in supervised, self-supervised, and Vision Transformer configurations. For example, a review and summary of the DR referable, non-referable, and proliferative classifications are provided. The study also covers the retinal fundus datasets for DR that are currently accessible and can be used for segmentation, classification, and detection tasks. Along with addressing several issues that require more research and analysis, the paper evaluates research gaps in the field of DR detection and categorization.

Keywords: Diabetes, retina, retinopathy, lesions, deep learning, diseases, biomedical imaging, diabetic retinopathy, diabetes mellitus, diabetic macular edema

INTRODUCTION

Although the cells in the pancreas cannot manufacture or secrete enough blood insulin, diabetes mellitus is a chronic illness characterized by elevated blood glucose levels [1]. Between 1980 and 2014, there were 422 million cases of diabetes worldwide, a sharp increase from 108 million cases in 1980 [2]. Diabetes has negative consequences on the liver, heart, kidneys, joints, eyes, and other human organs [1, 2]. For those under the age of 50, diabetes is the leading cause of blindness. Diabetic retinopathy (DR), a consequence of diabetes mellitus where glucose restricts blood vessels that supply the eye and produces swelling and leakage of blood or fluids that can cause serious eye damage, is directly caused by diabetes. The main cause of the harmful vision loss brought on by DR is central retinal edema. Our paper focuses on DR, glaucoma, and trachoma, which together account for an estimated 11.9 million cases of mild to severe visual impairment, according to the World Report on Vision [3]. Early identification is essential to prevent consequences from chronic diseases like diabetes. One possible side effect of

***Author for Correspondence**
Tshetiz Dahal
E-mail: dahaltshetiz21@gmail.com

[1]General Physician & Clinical Researcher, Department of General Medicine, Lugansk State Medical University, 16 Lypnia St. Rivne, Ukraine.
[2]Junior Resident, Department of Physiology, Rajendra Institute of Medical Sciences, Bariatu, Ranchi, Jharkhand, India.

Received Date: September 03, 2024
Accepted Date: September 09, 2024
Published Date: November 13, 2024

Citation: Tshetiz Dahal, Suyash Saurabh. The Analysis of Deep Learning-Based Methods for Identifying Diabetic Retinopathy. Research & Reviews: A Journal of Medical Science and Technology. 2024; 13(3): 15–31p.

DR is abnormal blood vessel formation in the retina, which can lead to retinal hemorrhage or scarring and ultimately blindness [3]. This may lead to a gradual loss of vision and, in more severe cases, blindness. DR accounts for 2.6% of blinding causes worldwide [4]. The longest duration of diabetes, elevated hemoglobin A1c, and elevated blood pressure levels are thought to be the main risk factors for developing diabetic kidney disease (DR) [5]. For diabetic patients, routine screening is essential to ensuring early detection of DR. Traditionally, the process of diagnosing DR entails a medical professional looking at retinal imaging to assess the appearance and shape of various lesions. Micro-aneurysms (MA), hemorrhages (HM), and soft and hard exudates (EX) are the four categories of lesions that are detected [6].

- Because of weakened artery walls, tiny red circular spots called MA can appear on the retina in the early stages of DR. Sharp edges that are no larger than 125 micrometers define the dots. Six subtypes of MA can be distinguished, although therapy for all of them is the same [7].
- Unlike MA, HM are identified by the presence of big patches on the retina with uneven edge widths up to 125 micrometers. The two types of hemorrhages – flame and blot – are distinguished by the depth and superficiality of their distinct patches [8].
- Hard EX, which appears as yellow patches on the retina because of plasma leakage, is a result. They have distinct borders and span the outer layers of the retina [1].
- Soft EX is visible as white ovals on the retina and is a result of nerve fiber swelling [6].

While the two forms of EX appear as brilliant lesions, MA, and HM typically appear as red lesions. The five stages of DR include no DR, mild DR, moderate DR, severe DR, and proliferative. Five potential stages of DR development are shown in Figure 1 [7].

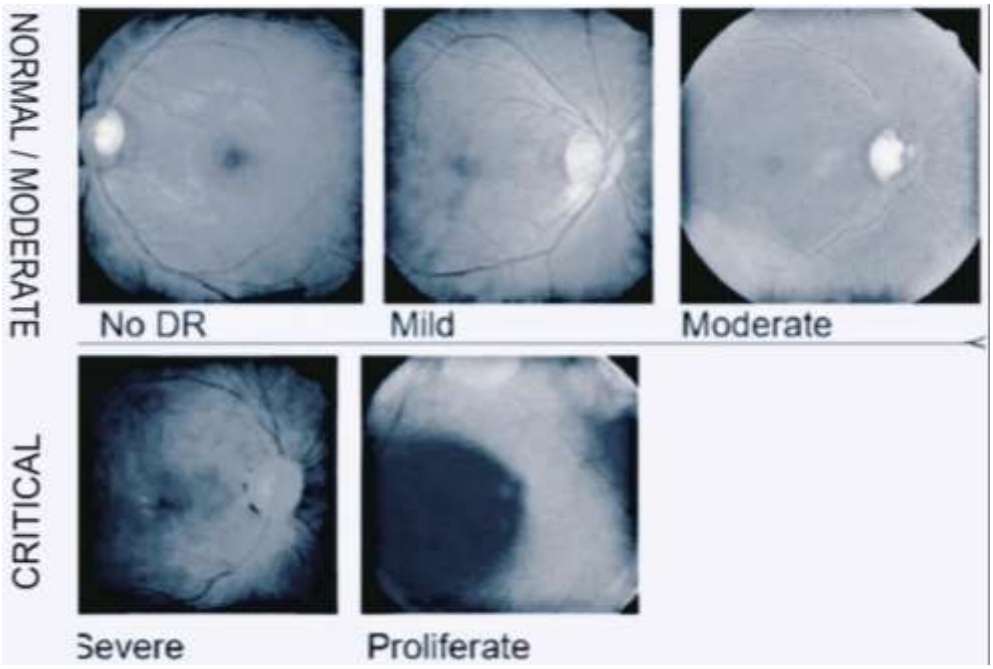

Figure 1. The 5 diabetic retinopathy stages, ranked by severity [7].

MA, intra-retinal HM, and venous beading – venous caliber alterations that alternate between zones of constriction and dilation – are among the retinal lesions that typically develops. In addition, it is recognized that the common types of lesions are retinal neovascularization [8], hard EX (lipid deposits), and intra-retinal microvascular abnormalities. Moreover, proliferative DR (PDR) and non-proliferative DR (NPDR) are the two primary stages of DR. Swelling may result from vascular leaking of fluid and circulating proteins into the retina due to damage to the retina's arteries. In this case, EX, HM, and MA may occur; this is referred to as NPDR. The diagnosis of non-diabetic PR is

based on the absence of neovascularization, which can include any of the frequent DR lesions discussed above. DR develops gradually with increasing severity through mild, moderate, and finally severe PDR that may endanger eyesight. By accurately classifying DR severity levels, high-risk patients can be identified and potentially mitigated through appropriate referrals, periodic examinations, and suitable therapy to maintain existing vision [9]. The latter stages of DR are referred to as PDR, which is an angiogenic retinal response. Angiogenesis is a physiological process in which preexisting blood vessels divide into new ones in PDR [10]. Neovascularization of the retina is generally understood to be the development of new blood vessels along the retina's vascular arcades [9]. Assessing manually for DR detection necessitates highly qualified practitioners. Furthermore, variations between and within grades might affect even highly qualified ophthalmologists. Therefore, such flaws may be mitigated by automated DR detection utilizing precise machine learning methods. Conventional retinal disease screening involves several scan steps, which are followed by filtration methods to reduce the subject samples. Among the scans carried out during the screening phase are optical coherence tomography (OCT) and spatial domain optical coherence tomography (SD-OCT). An ophthalmologist is then consulted for study of the fundus images that result. A substantial degree of intra-grader variability is usually present in this procedure; Figure 2 illustrates how DR grading can differ amongst specialists.

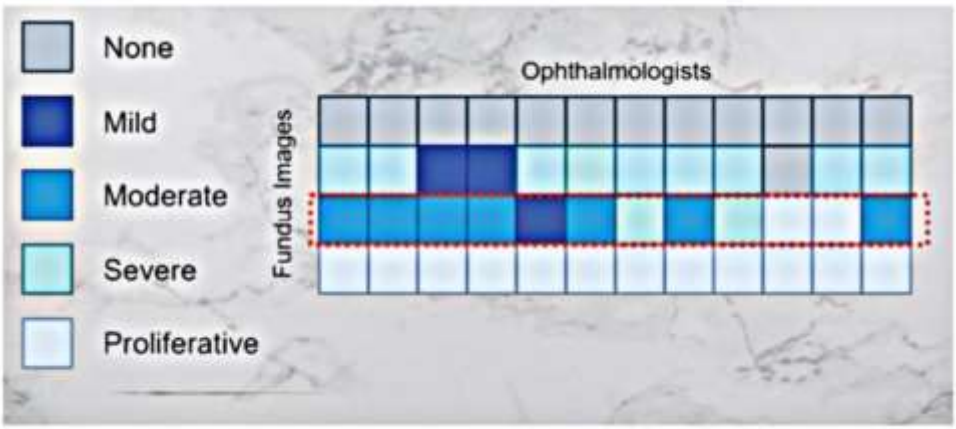

Figure 2. Inter-grader inconsistencies illustrated. In the highlighted red area, columns represent a single fundus image, and the rows represent the final grade provided by the ophthalmologist.

Several approaches have been developed to categorize OCT images. One method involves the use of local binary patterns (LBP), which have been refined over time and can be applied effectively for classifying OCT fundus images [11]. However, this method is not enough to differentiate between proliferative and NPDR cases. Retinal imaging uses infrared and multi-color laser to improve OCT outputs, allowing for considerably more accurate classification of fundus images. Although this method can identify lower-level anomalies like optic discs, it is still insufficient for accurately classifying DR cases

To further enhance model performance, significant attention has been directed towards developing more efficient image processing methods [12]. Exudate and hemorrhage traits are readily observable in computer-aided diagnosis (CAD). This makes it possible to separate low level, less significant lesions from mild and severe vascular abnormalities on fundus pictures by clustering them into proliferative and non-proliferative cases. Digital medicine has been made possible by CAD, which is one of the core diagnostic methods used in the medical field [13]. The several discernible DR lessons are depicted in Figure 3. We review the most recent works on DR categorization in this paper. This paper's focus is on how common deep learning (DL) approaches are for classifying DR data and how this affects the classification outcomes [14].

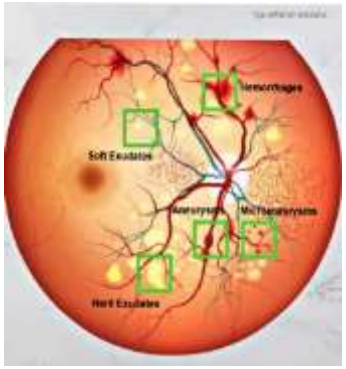

Figure 3. A fundoscopic illustration of the retina, showing micro-aneurysms, hemorrhages, and exudates.

EVALUATION OF SURVEY STUDIES IN THE LITERATURE

A recent survey explored various DR classification strategies, highlighting classical methods and placing particular emphasis on DL techniques. Additionally, another study examined different approaches to DR detection, focusing on advancements in the field [15]. gradually highlighted the shortcomings of the conventional methods in terms of learning more disease-related features. These methods include Adaboost, Random Forest, SVM, and other methods are commonly used for comparison in image analysis. Due to the poor contrast and quality of some publicly available datasets, these comparisons are often based on the quality of the fundus images [16] examined 33 publications in all that used DL for diagnosing diabetic kidney disease (DR) and stress the significance of ongoing advancements in DL models in light of the global rise in diabetes patients. The authors highlighted the importance of data augmentation during model training to reduce overfitting. After analyzing recent DL pipelines and machine learning approaches, several studies discussed various DR grading tasks, such as optic disc analysis, blood vessel detection, lesion identification, and overall grading [17].

In their discussion of new state-of-the-art CNN variations for DR classification, the authors highlighted the inconsistencies in evaluation metrics used in the literature to assess model performance [18]. They also provided a thorough explanation of how transformers operate across various medical imaging objectives, including segmentation, classification, detection, and reconstruction [19]**.** The analysis indicates that, with over 40 recent articles, transformer-based research for medical imaging peaked around December 2021. Additionally, according to the report, 27% of papers published between 2012 and 2015 used CNNs for segmentation tasks, while 73% of articles published in 2021 used vision transformers. This suggests that transformer-based methods are more in demand for segmentation jobs. Regarding retinal illnesses, [20] looks at a few visions transformer (ViT) papers that focus on lesion identification and DR grading and categorization.

Our review study provides insights into the DL-based DR categorization. We examine articles that discuss transformer-based methods and self-supervision as ways to lessen the dependency on massively annotated data. The literature's supervised, self-supervised, and transformer methodologies are covered in our main methodological review section [21].

COLLECTIONS OF DATA

Publicly available standard datasets, including DRIVE, EyePACS, APTOS, STARE, DIARETDB, HEIMED, ROC, Messidor, e-ophtha, DDR, and RFMiD, are the sources of retinal fundus images (RFI). These nine datasets are used to compare various DR classification methods. Private datasets are utilized in more focused research to enhance the accuracy of pretrained models. Typically, private datasets are small and originate from labs collaborating with researchers, and since these datasets are confidential, they are not publicly shared. An overview of all open-source DR datasets is provided. Most of the training data comes from EyePACS and Messidor-1 & 2. Fundus images were captured

using wide-lens Canon cameras with a 45°–50° field of view (FoV). The two largest datasets used are DDR, with 13,673 images, and EyePACS, with 88,702 images [22].

METHODOLOGY

DR classification can be divided into two categories: multi-class classification, which identifies the precise stage of DR, and binary classification, which seeks to identify the presence or absence of DR. As a result, other techniques centered on lesion-based classification were created. In the following sections of the paper, such classification tasks are reviewed under the context of supervised and self-supervised learning.

CONTROLLED PROCEDURES

Detection in Binary

To explore the use of CNNs for categorizing retinal fundus images with stochastic gradient descent as an optimizer, the authors conducted a study [23]. suggested a DL model. The authors have tested with various layers (ranging from 9 to 18) and kernel sizes (1 to 5). The resized fundus photos were (224, 224, 3). Using photos of both normal and DR from the Kaggle EyePACS [22] image dataset, the model was trained. To increase the diversity of images required to train the model and decrease overfitting, augmentations such as image rescaling, rotation, flipping, sharing, and translation were applied. There were about 200 testing photos, and 800 training images utilized in all. The retrieved features – hard EX, red lesions, MA, and blood vessel detection – formed the basis of the model. The paper's primary contribution was demonstrated by comparing the CNN-based model, both with and without augmentations, to the Gradient Boosting trees-based method combined with the extracted features (Hard exudates + GBM, Red lesions + GBM, MA + GBM, and blood vessel detection + GBM). The number of classes was set to two, with a maximum depth of six, using GBM hyperparameters. Because the Extreme Gradient Boosting method (XGBoost) outperformed other classical approaches in literature, it was employed. Utilizing the R programming language, the "MXNet" framework was applied Table 1.

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

Data Set	Img. Count	Img. Size (px)	Label Count	Train+Val+Test Size	DR Grading	Camera Used	No. of Studies
EyePACS 2015	88,702	433 × 289 to 5184 × 3456	{25,810, 2,443, 5,292, 873, 708}	{35,126, 53,576}	Yes	4 Cameras	3
APTOS 2019	3,660	Varies	–	–	Yes	Varies	2
Messidor	1,200	1440 × 960 to 2304 × 1536 (24-bit)	–	–	Yes	45° wide view	3
Messidor-2	1,748	1440 × 960 to 2304 × 1536 (24-bit)	–	–	Yes	45° wide view	3
IDRiD	516	4288 × 2848	{413, 103}	–	Yes	50° wide view Kowa VX-10	1
DRIVE	40	565 × 584 (24-bit)	{33, 7}	{20, 20}	No	45° wide view CANON CR5	1
DIARETDB1	89	1500 × 1152 (24-bit)	{5, 84}	{28, 61}	Yes	50° wide view	2
DIARETDB0	130	1500 × 1152 (24-bit)	{20, 110}	Varies	Yes	50° wide view	1
ODIR	10,000	Varies	{1,620, 8,380}	{9,000, 1,000}	Yes	45° wide view (42 cameras)	2
DDR	13,673	Varies	{6,266, 6,256}	{9,568, 4,105}	Yes	Topcon	1
RFMiD	3,200	2144 × 1424	–	{1,920, 1,280}	No	3 Cameras	1

To identify retinal fundus images as non-referable DR for stages 0 and 1 and referable DR for stages 2, 3, and 4, the authors conducted a study. trained three CNNs. For training and evaluation, private E-ophtha pictures and Kaggle DIARETDB1 [24] were utilized. After resizing the images, preprocessing steps included cropping them to 448 by 448, normalizing the pixels, and using a gaussian filter to reduce the Field of View (FOV) by 5%. Two networks comprising Team o_O solutions for the Kaggle competition plus a pretrained version of AlexNet comprised the model's architecture [25].

This approach categorized several cases, including soft and hard EX, HM, MA, and instances with no DR. The authors utilized learned weights in their analysis. efficiently integrated multiple DL model outputs using the Adaboost algorithm. Furthermore, class activation maps were produced by using the AdaBoost algorithm's results and learning weights in the same way. Inception V3 [26], Inception-Resnet-V2 [27], and Resnet152 [10] pretrained CNNs were used by the model to classify private datasets as referable or non-referable. Adam optimizers were employed in all three of their CNNs, and the AdaBoost algorithm was utilized to combine the output. Before being utilized for training, the dataset photos were preprocessed by resizing them to (520, 520, 3) pixels, then enhancing and augmenting them.

After cropping the imaging area, photos were standardized to 520 by 520 for preprocessing. Subsequently, the initial and altered fundus image undergo filtering and merging using weighted summation, yielding an improved image suitable for varying lighting scenarios. Prior to training, the following changes were applied: brightness, contrast, sharpness, translation, rotation, and mirroring. Three models were optimized using an Adam optimizer: the first model had an exponential decay rate of first-order moment estimation of 0.99 and the second model had a fixed learning rate of 0.001. By reducing the bias of each individual classifier, the model employs the local minimums produced by the three sub-models to discover the global minimum via the Adaboost algorithm. Three primary steps comprise implementation: distribution, initialization, iterative learning, and model combining.

For red lesion DR, enhanced image patches measuring 65 by 65 pixels were used. This was achieved by utilizing a pretrained VGG16 and CNN with five 2D convolutional layers, five max-pooling layers, and a fully connected layer. The training dataset for classifying red and non-red lesions came from a publicly available DR dataset, while the testing datasets included several other well-known DR datasets. Images were classified as either diabetic or non-diabetic based on the lesion probability map generated from the test cases.

MULTIPLE CLASSIFICATION

This section examines research that utilizes severity level classifications of normal, mild, moderate, severe, and proliferative to categorize the DR dataset. A convolutional neural network (CNN) model was employed to provide a method for diagnosing DR. Before being fed into their model, the preprocessed images were first normalized and then given a 299-pixel width for the diameter. Ten CNNs were trained as part of the model using a pretrained Inception-v3 architecture. Referable diabetic macular edema (DME), moderate or worse DR, severe or worse DR, and fully gradable were the five grades used in the classification. On their dataset, the approach was employed to identify DR.

The dataset comprised a total of 767 photos, which were divided into four classifications. The images underwent preprocessing, including cropping, resizing, histogram equalization, and adaptive histogram equalization. Augmentation techniques were applied to increase the dataset size, followed using a contrast stretching algorithm designed to enhance contrast in darker images. For the classification of DR, several pretrained CNN architectures, including ResNet50, InceptionV3, InceptionResNetV2, Xception, and DenseNets, were optimized. These CNNs served as the foundation for training new fully connected layers. The process included strong model integration, where the pre-trained CNN layers were fine-tuned for retraining. Additionally, a model was developed using a

modified R-FCN to identify different stages of DR for both the private dataset and another well-known dataset.

Their modifications involved incorporating a feature pyramid network and five region proposal networks into the R-FCN. Extensive data augmentation was applied to the training images, particularly those from the private dataset. The authors classified images into referable and non-referable categories for the Messidor dataset. While four attention modules and ResNet50 were used to classify images from the publicly available IDRiD dataset into five classes (class 0 to class 4). The first attention modules received inputs in the form of the features that ResNet50 extracted. The first two attention modules contain average pooling, max-pooling, multiplication, concatenation, 2D convolution, and fully connected layers. The last two attention modules, on the other hand, only have fully connected layers and multiplication.

The preprocessing of the images included scaling, normalization, and augmentation. Bi-channel neural networks were utilized in the analysis. to extract fundus components by channel. Unsharp masking (UM), a traditional sharp enhancement method, was then used to increase detail. This method made use of 21,123 RGB fundus picture sizes that were chosen from the Kaggle Diabetic Retinopathy dataset [22]. We downsized the pictures to (100, 100, 3). Then, 33,000 photos are produced for the experiments by flipping and rotating. To create the 30,000 fundus images for the training set, 15,000 randomly selected samples are taken from the first group of grade 0 fundus images and another 15,000 samples are taken from the second group of grades 1–4 fundus photos. Similarly, 3,000 photos are selected for the test set. The gray level entropy images and the green component of the fundus are used to train the bi-channel CNN for feature learning of referable DR. The images are first preprocessed using UM to improve the detection of DR, particularly the referable type. The green component of the retinal image and gray-level high-frequency portions are amplified using the Unsharp Marking approach. Four convolutional layers with five-by-five kernels and feature map sizes/number of filters of 32, 64, and 128 are employed per channel, accordingly. Maximum pooling rectified linear unit activation function (ReLU), and dropout layers – dropout set to 0.3 – are employed for each layer. Flattening for the two channels comes next, and the fully linked layers linkage is utilized to statistically classify referable DR.

To categorize DR classes, a multi-task learning strategy that combines a deep CNN architecture with a tiny decoder, specifically a head and a feature extractor was used. The CNN was pretrained using the Kaggle EyePACs dataset [22]. Two more datasets that were merged to create the training set were the MESSIDOR dataset [28], which comprises 1200 fundus photos, and the IDRiD dataset, which contains 413 fundus shots. Optic distortion, grid distortion, piecewise affine transform, flipping the images horizontally or vertically, random rotation, random shift, random scale, shifting the RGB values, random brightness and contrast, additive Gaussian noise, blur, sharpening, embossing, random gamma, and cutout are some of the changes that were made to the images.

The encoder in the model is initialized using CNNs that have already been trained using ImageNet. They employ three decoders, each of which has been trained to use the extracted features to accomplish a different job utilizing the CNN backbone's classification, regression, and ordinal regression heads. On the other hand, the output of the classification head is a one-hot encoded vector, where each stage's existence is represented by a value of 1. The output of the regression head represents the various phases of the disease as a rounded real number between 0 and 4.5. Regarding the ordinal regression head, each data point inside a category is assumed to fall into every other category, thereby forecasting every category up to the target category. The total prediction is obtained by fitting a linear regression model to the outputs of three heads using an ensemble of three heads. This ensemble is predicated on the disease's sequential nature, which has been assessed using the Kaggle APTOS 2019 dataset [23].

SELF-DIRECTED TECHNIQUES

Not all problems lend themselves well to supervised learning techniques, particularly when the data is noisy. Self-supervised learning (SSL) techniques can be utilized in conjunction with supervised techniques or as a perfect substitute for them. SSL techniques can handle cross-domain inputs and are less susceptible to inductive bias. The issue with SSL techniques is that their effectiveness depends on large amounts of data. This is problematic because training the model takes longer than the more data you have.

IDENTIFICATION IN BINARY

A Self-Supervised Fuzzy Clustering Network (SFCN) was developed to enhance the analysis of the data. It consists of three primary modules: a reconstruction module, a fuzzy clustering module for self-supervision, and a feature learning module for unlabeled retinal fundus pictures. Given an input fundus picture, the feature learning module is first composed of convolutional layers for feature representation extraction and then de-convolutional layers for retinal image reconstruction. This module contains all the information required to rebuild the retinal pictures. The fuzzy self-supervision module then uses the predictions of the fuzzy clustering module technique to give the feature learning module training supervision. Fuzzy clustering is used to infer the connection in unlabeled retinal pictures using these self-supervised models, with probability belonging to each respective cluster. ResNet50 [25] was used to create the feature learning module, which consists of a fully connected layer at the end, four residual blocks, and one convolutional layer. The model architecture suggested by Johnson and Fei-Fei [26] constituting a residual block with two stride-1/2 de-convolutions for up-sampling purposes with an attached instance normalization was utilized to create the image decoder, which was then achieved using the Cyclic GAN [22] vanilla image decoder. The fuzzy C-means clustering output is fully incorporated in the feature module generated from the convolution layer stack in the fuzzy self-supervision module. Two fully connected layers are utilized to output the predictions after the feature learning module. Images are downsized to (23, 24) with a batch size of 32 during training and testing. Utilizing an SGD optimizer, their initial learning rate of 0.001 decayed to 0 by the end of the 300 training epochs.

MULTIPLE CLASSIFICATION

A novel category attention block (CAB) was introduced to improve model performance. so that features based on regions may be tested for each DR grade. DR multi-class classification using this network is frequently employed to reduce the DR grade imbalance in distribution found in most publically accessible datasets, including DDR, Messidor, and EyePACS. The use of category attention acts as a complement to spatial and channel attentions, allowing the CAB to be integrated with various non-category-centric blocks to enhance multi-class classification, specifically for DR grading. This model is inspired by previous work in the field. and blends the previously stated Cabinet with GABNet, i.e., CABNet is suggested for DR grading in GAB and CAB. While disregarding qualities like color and texture, GAB can learn global class-eccentric traits. In concert, CABNet collects fine-grained lesions features to address the issue of unequal data distribution. The CABNet module is composed of four components: a classifier, a global attention block (GAB), a CAB, and the backbone. The GAB and CAB, for which the entire CABNet training is undertaken, make up the attention module.

The CABNet receives input fundus images, and the backbone network is only utilized to gather and extract feature maps globally. Since the model is adaptable, any CNN architecture can be utilized as the foundation, with features taken from the input fundus images' exceptionally rich semantic properties at the final convolutional layer. Next, using the feature map that was previously extracted from the backbone, a 1 by 1 convolutional layer is fed in for input channel reduction. This layer is then supplied as input to GAB, and the output of the spatial attention is fed as input to CAB, before being fed into a classifier for DR grading. The backbone network of the base CABNet model has been pretrained using the ImageNet dataset. Random rotation, random vertical flips, and random horizontal flips using images of size (512, 512, 3) are among the data transformations used. Based on validation

loss, the learning rate was first set at 0.005 and then gradually decreased by a factor of 0.8. Using a cross-entropy loss function and an Adam optimizer, training is carried out for 70 epochs. Various backbone models were trained, and the base model is the best-performing model with the lowest validation loss. 16 was chosen as the batch size.

A Graph Convolutional Network (GCN)-based module, referred to as MCG-Net, was introduced to enhance data analysis. to facilitate the effective feature extraction of fundus image lesions for multiclass classification. This enhanced lesions classification. A GCN is utilized in place of a fully connected layer in an augmentation module of the previously introduced MCGS-Net based on SSL to better capture the correlation of fundus pictures as a classifier and boost generalization. The CNN's capacity for generalization improves when self-supervised learning is used.

TRANSFORMER TECHNIQUES

The attention method employed for text-based data was initially applied to images by [28] with the ViT proposal. Image inputs in ViT are divided into patches of (16,16), projected onto positional marker embeddings, and subsequently routed to the transformer encoder layers. The procedure in the encoder is akin to that of text-based data; the patch embedding flow via a multi-head self-attention block that learns and concatenates the image's local and global dependencies. Following normalization for generalizability, the output of the self-attention layer is routed to the final MLP head, where classification takes place. Numerous studies have made use of this function to classify DR. By introducing a novel lesion-aware transformer using a unified model with 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**,** helped close the gap between DR grading and lesion discovery. The fully connected layer and the global average pooling (GAP) layers are eliminated for feature extraction, which is the primary function of the model's ResNet50 backbone. After resizing the images to (512, 512) pixels, augmentations are applied to increase the number of training images and reduce overfitting. These augmentations include random cropping, vertical and horizontal flips, and color jitter. Datasets from multiple sources, including well-known DR datasets, are then used for testing. To synthesize fluorescein angiography (FA) from fundus images – an exogenous dye injected into the bloodstream to visualize the retinal vascular structure and highlight retinal degeneration – a novel state-of-the-art conditional generative adversarial network (GAN) is proposed. The model employs a semi-supervised training GAN configuration with varying losses and corresponding weights to provide a non-invasive alternative while still benefiting from the dye. This ViT-based GAN utilizes transformer encoder blocks for the discriminators and incorporates residual, spatial feature fusion, as well as upsampling and downsampling modules for the generator [29].

The model creates FA pictures using training images of normal and diseased funds. To extract 50 photos with a crop size overlap of (512,512) from each corresponding set, the original images of size (576,720) are utilized for training. A total of 850 photos are extracted for image synthesis training. FA images only have one channel, whereas fundus images have three channels (RGB). The private training and testing datasets are divided into abnormal and normal classes, and the supervised classification training portion uses the annotation for each class. The private dataset has 17 photos in total, 10 of which are aberrant and 7 of which are of normal patients. Due to cropping, as previously indicated, this then expands to 500 photos for abnormal class and 350 for normal class. By combining multiple rectangular patches with an effective attention structure concentrated on eye regions with lesions (abnormal images), Papadopoulos presented a transformer-based method for the independent extraction of necessary local information for classifying images. The model also generates heatmaps using the attention mechanism. The image preprocessing procedure involved several steps, including the Hough transform, randomly resized cropping, removing the local color average of the image to account for varying lighting conditions, and zeroing out the outer 5 percent of the retina disk. The model was trained using the Kaggle EyePACs dataset, and tests were conducted on the Messidor-2, Indriid, and Kaggle EyePACs datasets.

EVALUATION OF THE OUTCOMES

The most innovative investigations are highlighted in this section along with their published findings. The goal is to evaluate, compare, and examine the scalability and generalizability of the best-performing approaches. Table 2 and Table 3 present all the findings.

Table 2. Evaluation of metrics.

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

Table 3: DL Methods for DR detection & classification.

Lesion Det.	Classifier	Arch.	SSL	Dataset	Acc	Recall (SN)	SP	AUC	F1	Kappa
No	Binary	CNN	No	EyePACS	94.50%	–	–	95.40%	–	–
No	Binary	CNN	No	EyePACS, DIARETDB1	–	–	–	94.90%	–	–
No	Binary	Inception-V3 & ResNet152	No	Private dataset	88.21%	–	–	94.60%	–	–
Yes	Binary	VGG16	No	DIARETDB1	94%	–	–	91.20%	–	–
No	Multi	Inception-V3	No	EyePACS	–	97.50%	93%	–	–	–
No	Multi	Multi-CNN	No	Private dataset (13,767)	96.50%	98.10%	98.90%	98.62% (Ensemble)	95.42%	–
Yes	Multi	CNN	No	Private dataset (9,194), Messidor (1,200)	92.95%	99.39%	99.93%	–	–	–
Yes	Multi	ResNet50	No	Messidor, IDRiD	92.60%	92%	–	96.30%	91.20%	–
Yes	Binary	Bi-channel CNN	No	EyePACS	87.37%	76.93%	93.57%	93%	–	–
Yes	Binary	CNN (ResNet-101)	No	DRIVE	95.10%	79.30%	97.40%	97.32%	–	–
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										
Yes	Binary	CNN + Fuzzy C-Means	Yes	DRIVE, Messidor, DIARETDB1	95.10%	–	–	–	–	–
Yes	Multi	CNN + CAB & GAB	Yes	EyePACS	84.00%	–	–	–	85.50%	–
Yes	Multi	CNN + GSSL & GCN	Yes	ODIR, SSL, GTest	–	–	–	–	86.56%	38.77%
Transformers										
Yes	Multi	CNN + ViT	No	Messidor-1&2, EyePACS	–	–	98.70%	–	88.40%	–
No	Binary	GAN + ViT	No	Private dataset	85.70%	83.30%	90.00%	–	–	–
Yes	Binary	ResNet18+	No	Messidor 2,	99%	92.00%	99.00%	–	–	–

		Attention (MIL+Ensemble)		EyePACS						
No	Binary + Multi	ViT (MIL)	No	APTOS, RFMiD	85.5% (Binary), 93.7% (Multi)	–	97.9% (Multi)	85.3% (Multi)	92.0% (Multi)	–

Note: Highest scores reported.

METRICS FOR EVALUATION

The primary metrics for assessing the validity of the results for DR detection and classification are the AUC, F1, and Kappa scores; other metrics, such accuracy and recall, can be viewed as support metrics. This is a result of each dataset's uneven data distribution. Table 2 presents the formulas used to determine the metrics for a few use cases.

OUTCOMES UNDER OBSERVATION

The Categorization in Binary

A model that produces entropies of every fundus picture is presented, which enables it to further highlight the edges of lesions and produce areas of interest for the feature extractor during binary classification. With this approach, it obtains an F1 score of approximately 81.8% and an accuracy of 87.37%. Ignoring instabilities in the fundus data, the AUC of the ROC curve serves as a fundamental performance metric in this work. While some attributes are extracted from lesions in multiple experiments, the final output is unable to determine the type of lesion. This process is further developed by identifying the DR stage as one of the five stages: mild, moderate, severe, or proliferative. Similarly, another model achieved multi-class accuracy, specificity, and sensitivity of 96.5%, 98.9%, and 98.1%, respectively, and was able to distinguish between the various severity levels. However, the results were limited to the private data set used for assessment, and this model is unable to identify retinal fundus abnormalities. Although a unique model, RFA-BNET, notable for its methodology, is based on ResNet-101, it obtains a comparatively lower accuracy rate by aggregating features from multiple rounds through ResNet-101. A recall of 79.3% and accuracy of 95.1% were reported with this approach**.**

Multiple Classification

By utilizing an ensemble of 20 distinct models and a trimmed mean of 200 five-class predictions for each fundus image, the study achieved optimal results through transfer learning in a 3-headed ensemble CNN architecture, incorporating classification, ordinal, and regression approaches. With training time augmentations (TTA), the model achieved an accuracy of 99.3%. Through binary classification screening, the model's quality was assessed, resulting in an F1 score of 99.3%. When compared to studies utilizing a single ResNet50 layer, the findings are comparable to what can be attained with more complex models, such as those incorporating an ensemble of pretrained networks, which have evolved into comprehensive frameworks for DR detection, demonstrating the potential of high-grade ensemble networks.

For the Messidor dataset, sensitivity, AUC, and accuracy were reported as 92%, 96.3%, and 92.6%, respectively. For the IDRiD dataset, the achieved accuracy was approximately 65.1%; however, precision and F1 scores were not reported. Another study presents a VGG16 network that achieves a sensitivity of 0.94 and an AUC of 0.912 for the Messidor dataset. However, this model is limited to detecting DR without providing any indication of severity, aiming to reduce the number of trainable parameters. Similarly, no F1 score or accuracy was mentioned.

RESULTS OF SELF-SUPERVISION

Classification in Binary

The SFCN model, trained on 25 DRIVE pictures. It succeeded in obtaining an accuracy of 81.7%. By contrast, RFA-BNET obtains a 95.1% accuracy rate using 20 images that have undergone

significant augmentations. There were no documented tests demonstrating its generalizability to other sets, such as EYEPACs.

Classification Using Multiple Classes

After training for 70 epochs on the EYEPACS dataset, CABNet was able to generalize to Messidor and obtain a very competitive kappa of 85.5% and accuracy of 84%, respectively. It is simpler to understand how the model functions and which regions of the fundus images contribute to the embeddings thanks to CABNet's specially designed attention block. With this architecture, CABNet has an edge in terms of customizing the attention blocks and adds to its flexibility.

Lin and others on the other hand, MCGG-Net obtains an F1 and kappa of 86.56% and 38.77% on the GTest dataset, respectively, after training with 60 epochs on 9000 ODIR pictures. These findings demonstrate that the model, which utilizes only the embeddings of an image with few or no labels, may generalize well to additional fundus images from a completely different dataset. In comparison, a top-performing multi-class supervised model achieved a kappa of 96.9% and an F1 score of 84%. This three-headed CNN makes substantial use of augmentations through a complicated ensemble training method. Whereas CABNet employs both flips and rotations, MCG-Net only augmentations flip once. Due to their high efficiency in the data preprocessing phase, SSL models can generalize much better than supervised models and can be adjusted and adjusted to new fundus images more quickly. This is advantageous when conducting research on a variety of datasets is necessary.

TRANSFORMER OUTCOMES

Binary Arrangement

The GAN evaluation system known as VTGAN, initially introduced in previous studies, utilizes two assessments. The architectural performance is evaluated qualitatively using two metrics: the Kernel Inception Distance (KID), which examines visual characteristics and structural similarities to the original fundus counterpart, and the Frechet Inception Distance (FID), which measures the image quality of the GAN model. As compared to SoTA, the results show at least 30% higher FID and KID scores for models like A2GA and StarGAN-v2. Furthermore, VTGAN outperforms SoTA by 5% in the converted images. Overall, there is an average quantitative correctness of 78% on out-of-distribution transformed images compared to 85.7% on in-distribution funds, and an average qualitative precision score of 45.9% overall.

Using multiple instance learning (MIL) in its attention mechanism, another MIL study treats patches of DR lesions as a bag of features, keeping only pertinent information for the classifier to work with. By removing the black box, this method made it possible to produce excellent attention maps that emphasize lesions that have been identified. The MIL model obtained 99% AUC on Messidor-2 [28] and 95.7% AUC on Kaggle EYEPACS [22] using random patch selection. Additionally, employing a lesion classification approach, the paper investigates how lesions affect attention weight. It was concluded that larger attention weights were produced by smaller lesions with more features or variations. For all lesions (MA, HM, and EX), an AUC of 80% was attained.

CLASSIFICATION USING MULTIPLE CATEGORIES

Another comparable MIL model that can categorize the degree of DR sickness is MIL-VT. Except for the addition of a MIL embedding layer that combines patches based on features and attention prior to delivering them to the MIL classifier (MIL Head), the methodology utilized in MIL-VT is nearly comparable to the standard ViT approach for generic pictures. MIL-VT obtained 85.5% accuracy in DR grading and 94.4% F1 in DR disease categorization on APTOS [23]. GREEN-SE-ResNext50 attained an accuracy of 85.7% and an F1 score of 85.3% in comparison to SoTA. DR grading is carried out via lesion aware transformer (LAT), which uses attention blocks instead of complete transformer topologies. The concept is that contextual aspects of lesions are encoded and output by a self-attention layer, which is subsequently combined into the final result, obtained a DR grading accuracy of 96.3% on normal fundus and 98.7% on fundus with detected lesions (referrals) by employing a technique called lesion significance learning. This performs better than SoTA, including

Semi+Adv and CANet. The ablation study designed to evaluate the five network components – pixel relation encoder (P), self-attention layer (S), cross-attention layer (C), region diversity mechanism (D), and global consistency loss (G) – is what distinguishes the evaluation from others.

DISCUSSION

This work examines eleven supervised, three self-supervised, and four transformer publications by examining the methodologies and outcomes of each approach. This paper's major goal is to evaluate DR grading and classification techniques from a qualitative perspective. By doing so, it will essentially enable future research projects to be aware of the developments in the DR sector. Of the evaluated papers, 46% classify DR by four severity categories, while 54% employ binary classification for DR detection in the supervised approaches. About 33% of the self-supervised techniques use binary classification, while 67% use multi-class detection. Regarding datasets, a few studies attempted to employ privately constructed sets for complete control as demonstrated by these efforts typically produced improved findings. and however, because of different data distributions, no association can be inferred. Nearly 42.8% of the research employs a single dataset for model training, whereas about 57.2% of the studies train their models on several datasets. A summary of the gathered data is shown in Figure 4. Supervised models are reliable and excellent for certain applications. Frequently overlooked in most articles is the enormous amount of time required to train new models each time. The effort of annotating and structuring data to make it "model ready" almost seems counterproductive for mission essential models that are continuously inundated with new data. In the context of supervised learning, ensemble learning pipelines in DL should be encouraged as they excel at managing features in variable data distributions.

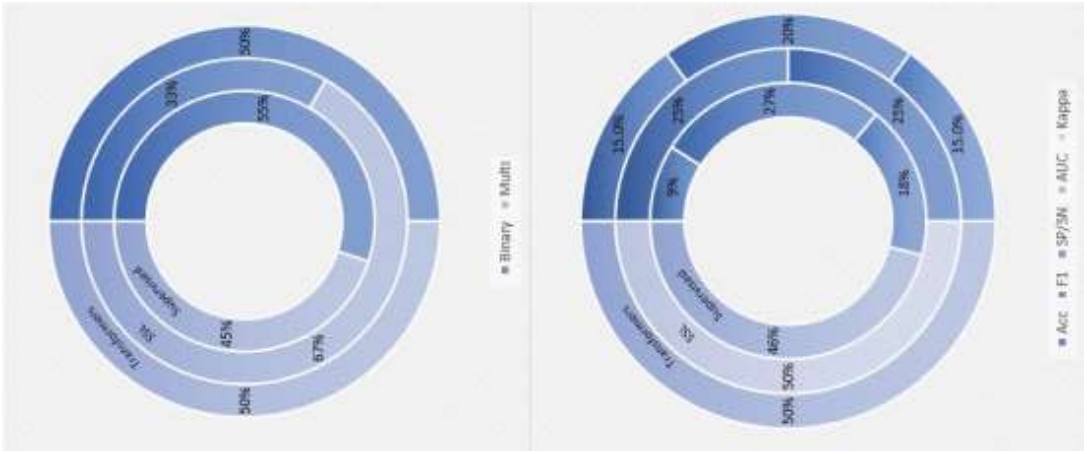

Figure 4. The classification segment distribution is displayed in the top chart for four transformer experiments (outside circle), three self-supervised studies (middle circle), and eleven supervised studies (inside circle). The bottom chart displays the distribution of assessment measure usage across four transformer studies (outside circle), three self-supervised studies (middle circle), and eleven supervised studies (inside circle).

Self-supervised models have superior performance when it comes to fine-tuning on fresh datasets. According to all three publications, the model might generalize to a larger range of datasets, including DRIVE and GTest, if it was trained on a very big dataset, like Messidor or EYEPACS. The results that were given demonstrated this to be accurate. But one must comprehend how SSL models read the data that is fed into them to make sense of them. The ability of SFCN to separate normal and abnormal images is interpreted by t-SNE plots. Attention maps are used by A. To illustrate the qualities that the model prioritizes. The heatmaps show how the attention block of CABNet aids in limiting the characteristics that are chosen. In essence, this can lower the size of embeddings required for generalization. Studies demonstrate the effectiveness of self-supervised approaches over supervised techniques, but they do not demonstrate how SSL approaches can ultimately be less

susceptible to inductive bias. When choosing a model for mission-critical applications, this kind of advantage is especially important because SSL methods are known to handle cross-domain inputs well. Furthermore, the resilience of SSL approaches when applied to small-scale datasets is not addressed in the articles. Due to the scarcity of publicly available datasets, DR screening is still an open problem. Although most current DL innovations acquire encouraging classification scores, some are still unable to differentiate between lesions that are impacted. Some approaches simply overlook the five stages of DR, which are thought to be essential for assessing the illness's severity. This kind of technical disparity highlights another obstructive issue. Researchers in the field may still disagree on the validity of the findings, as seen by the lack of conventional protocols that settle on a set of DR phases. However, most of the results are never considered final and are just useful for diagnosis.

Future research should concentrate on utilizing SSL techniques to create fresh fundus images based on the attributes that are acquired through generative networks, in addition to allowing for generalization. Combining current networks with GAN and variational auto-encoders (VAE) allows for the synthesis of a wide variety of augmented fundus pictures that may be made available for training. For instance, DALL-E model can produce images from text; this means that a model of this kind might produce a sizable collection of DR fundus images that could be used for testing and training. When large-scale DR sets are available, another approach would be to use self-supervised vision transformers, like DINO, which suggested being used to encode superior features. Transformers are resistant to saturation with larger sets and different data distributions since research has revealed a positive association between accuracy and the number of trainable parameters. The growing complexity of CNN layers as filter sizes grow is what vision transformers would address. Because they might not maintain visual elements across the network, CNNs are unable to obtain a comprehensive knowledge of the image. Concurrently, vision transformers leverage their attention mechanisms to identify patterns within flattened feature sequences. According to current research, attention mechanisms have significantly altered how people understand and contextualize visuals. In the field of medical imaging, the transformer-based models surveyed exhibit encouraging results for both binary and multi-class categorization of DR illness. A noteworthy advancement observed in these transformer investigations is the capacity to differentiate between smaller lesions with greater precision, which enhances the explainability of the resulting categorization. Even though the results demonstrate improvements over CNNs and performance that is sufficient, more research is needed to benchmark the models in real-world deployment scenarios. Once implemented in the actual world, vision models typically don't work as intended.

One example that appears to function quite well in theory is VTGAN, however it might not capture some aberrations that are not normally synthesized in the process of creating fundus images. For example, warping, blur, and noise might all occur in real fundus scans by coincidence, but they are not taken into consideration collectively in the qualitative evaluation because of network design constraints. Conversely, transformer-based research brought in more useful white-box strategies, like various obligation and assessment methods that focus on the explainability and rationale of outcomes. The models incorporate context more effectively and generate attention maps that highlight discovered lesions using attention maps and feature enrichment approaches. In the real world of business, explainable structures are thought to be more realistic. Every choice made in the medical field must be supported by knowledge, investigation, and scientific evidence. Integrating medical imaging models into the end-to-end operational pipelines of numerous institutions and research institutes requires interpretable design.

CONCLUSION

Even though there is no cure for DR, it is crucial to identify it early to stop more harm from happening. For instance, early signs of DR are nearly always present in NPDR stages and being able to identify and categorize those stages with the use of an appropriate evaluation technique may be the difference between saving one's vision. A significant amount of the work in this review paper is

devoted to the investigation of EX, MA, and HM. Multiple studies' results indicate a promising overall categorization performance with an accurate average of roughly 91%. Modern screening systems might use these DL-based methods to improve and categorize the DR stage by applying lesion detection strategies to a variety of fundus pictures. The primary concern discussed in the reviewed research is the need for manual diagnosis following screening, which is usually a time-consuming procedure subject to ophthalmologists' prejudice. Additionally, fundus image variations that are useful for evaluating indications are limited by dataset restrictions.

The analysis of retinal scans has grown more rapid, inclusive, and generalizable due to the effectiveness of DL algorithms. However, the metrics employed to evaluate the outcomes, and their corresponding datasets continue to be biased and uneven among various studies. In the end, categorizing DR is important, but comprehending its numerous origins might also be a worthwhile study project. For example, particular alterations in lesions and other subtle clues may suggest the development of DR. Since the presence of DME is strongly suggestive of the development of DR, further research avenues could entail investigating DME. These developments allow for the generalization of DL-based models and the evaluation of a greater variety of symptoms and signs, which may aid in the investigation of the etiology of retinal-based disorders.

Transformers have introduced more interpretable techniques that help address the limitations of non-generalizability. By patching and embedding fundus images using various context enrichment methods, hidden signs can now be detected with greater accuracy. Notably, only two studies in SSL and Supervised techniques stood out for their interpretability: one utilizing CABNet and another employing a CNN Ensemble model with SHAP analysis. The distribution of studies based on network interpretability and approach is illustrated, highlighting how transformer-based models enhance the interpretability of research and future work.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Funding Statement

This study did not receive any funds.

REFERENCES

1. Amin J, Sharif M, Yasmin M. A review on recent developments for detection of diabetic retinopathy. Scientifica (Cairo). 2016;2016: 6838976. doi: 10.1155/2016/6838976.
2. Kharroubi T, Darwish HM. Diabetes mellitus: The epidemic of the century. World J Diabetes. 2015;6(6):850–867. doi: 10.4239/wjd.v6.i6.850.
3. Mamtora S, Wong Y, Bell D, Sandinha T. Bilateral birdshot retinochoroiditis and retinal astrocytoma. Case Rep Ophthalmol Med. 2017;2017: 6586157. doi: 10.1155/2017/6586157.
4. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35(3):556–564. doi: 10.2337/dc11-1909.
5. Dubow M, Pinhas A, Shah N, Cooper FR, Gan A, Gentile CR, et al. Classification of human retinal microaneurysms using adaptive optics scanning light ophthalmoscope fluorescein angiography. Invest Ophthalmol Vis Sci. 2014;55(3):1299–1309. doi: 10.1167/iovs.13-13122.
6. Vora P, Shrestha S. Detecting diabetic retinopathy using embedded computer vision. Appl Sci. 2020;10(20):7274. doi: 10.3390/app10207274.
7. Murugesan N, Üstunkaya T, Feener EP. Thrombosis and hemorrhage in diabetic retinopathy: A perspective from an inflammatory standpoint. Semin Thromb Hemost. 2015;41(6):659–664. doi: 10.1055/s-0035-1556731.

8. Wong YT, Sun J, Kawasaki R, Ruamviboonsuk P, Gupta N, Lansingh VC, et al. Guidelines on diabetic eye care: The International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018;125(10):1608–1622. doi: 10.1016/j.ophtha.2018.04.007.
9. Birbrair A, Zhang T, Wang ZM, Messi ML, Mintz A, Delbono O. Pericytes at the intersection between tissue regeneration and pathology. Clin Sci (Lond). 2015;128(2):81–93. doi: 10.1042/CS20140278.
10. Silva S, Bouwmans T, Frélicot C. An eXtended center-symmetric local binary pattern for background modeling and subtraction in videos. Proc 10th Int Conf Comput Vis Theory Appl. 2015;395–402. doi: 10.5220/0005266303950402.
11. Al Hazaimeh M, Nahar KMO, Al Naami B, Gharaibeh N. An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images. Int J Signal Imag Syst Eng. 2018;11(4):206. doi: 10.1504/IJSISE.2018.10015063.
12. Fujita H, Uchiyama Y, Nakagawa T, Fukuoka D, Hatanaka Y, Hara T, et al. Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs. Comput Methods Programs Biomed. 2008;92(3):238–248. doi: 10.1016/j.cmpb.2008.04.003.
13. Attia Z, Akhtar S, Akrouf S, Maza S. A survey on machine and deep learning for detection of diabetic retinopathy. ICTACT J Image Video Process. 2020;11(2):2337–2344. doi: 10.21917/ijivp.2020.0332.
14. Gupta R, Chhikara R. Diabetic retinopathy: Present and past. Proc Comput Sci. 2018;132:1432–1440. doi: 10.1016/j.procs.2018.05.074.
15. Alyoubi WL, Shalash WM, Abulkhair MF. Diabetic retinopathy detection through deep learning techniques: A review. Informat Med Unlocked. 2020;20:100377. doi: 10.1016/j.imu.2020.100377.
16. Stolte S, Fang R. A survey on medical image analysis in diabetic retinopathy. Med Image Anal. 2020;64:101742. doi: 10.1016/j.media.2020.101742.
17. Asiri N, Hussain M, Al Adel F, Alzaidi N. Deep learning-based computer-aided diagnosis systems for diabetic retinopathy: A survey. Artif Intell Med. 2019;99:101701. doi: 10.1016/j.artmed.2019.07.009.
18. Valarmathi S, Vijayabhanu R. A survey on diabetic retinopathy disease detection and classification using deep learning techniques. Proc 7th Int Conf Bio Signals Images Instrum (ICBSII). 2021. pp. 1–4. doi: 10.1109/ICBSII51839.2021.9445163.
19. Shamshad S, Khan S, Zamir SW, Khan MH, Hayat M, Khan S, et al. Transformers in medical imaging: A survey. Med Image Anal. 2023;88:102802. doi: 10.1016/j.media.2023.102802.
20. Asad H, Azar AT, El-Bendary N, Hassaanien AE. Ant colony-based feature selection heuristics for retinal vessel segmentation. 2014. arXiv:1403.1735 [cs.CV].
21. Diabetic Retinopathy Detection EYEPACS Dataset. San Francisco, CA, USA. 2015.
22. APTOS. Atlanta, GA, USA. 2018 Jun. Available at https://www.kaggle.com/c/aptos2019-blindness-detection
23. The STARE Project. San Diego, CA, USA. 2004.
24. Kauppi T, Kalesnykiene V, Kamarainen JK, Lensu L, Sorri I, Uusitalo H, et al. DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms. 2007;61–65. doi: 10.5244/C.21.15.
25. Giancardo L, Meriaudeau F, Karnowski PT, Li Y, Garg S, Tobin WK, et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med Image Anal. 2012;16(1):216–226. doi: 10.1016/j.media.2011.07.004.
26. Niemeijer M, van Ginneken B, Cree MJ, Mizutani A, Quellec G, Sánchez CI, et al. Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imaging. 2010;29(1):185–195. doi: 10.1109/TMI.2009.2033909.
27. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis. 2020;59:101561. doi: 10.1016/j.media.2019.101561.

28. Decenciere E, Cazuguel G, Zhang X, Thibault G, Klein JC, Meyer F, et al. TeleOphta: Machine learning and image processing methods for teleophthalmology. IRBM. 2013;34(2):196–203. doi: 10.1016/j.irbm.2013.01.010.
29. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci. 2019;501:511–522. doi: 10.1016/J.INS.2019.06.011.

