Classification of Diabetic Retinopathy (DR) using ECA Attention Mechanism Deep learning Networks

Doha S. Salem[1], [*], M.Sami soliman[1], Ahmed Donkol[1], Gamal M. Dousoky[1], [2]

[1]*CCE Dept., faculty of Engineering, Nahda University, Beni-Suef, Egypt*

[2]*Electrical Engineering Dept., Minia University, Minia, Egypt*

Email: dohaa.salah@nub.edu.eg

A R T I C L E   I N F O

A B S T R A C T

Diabetic retinopathy (DR) is the most frequent eye condition among diabetics and a leading cause of blindness. Effective management of the disease requires regular fundus photography screening and prompt action. A computer-aided and entirely automated diagnosis of DR has attracted attention due to the increasing number of diabetic patients and the extensive screening they need. In recent years, advanced deep neural networks have been widely used in various fields. We propose to early detect the DR using the Efficient channel attention (ECA) model. For DR color medical picture severity detection, a deep convolutional neural network model called ECA-Resnet101 (ERNet), ECA-VGG19 (EVNet), and ECA-InceptionV4 (EIANet) have been constructed using a flexible one-dimensional convolution kernel size approach dependent on the feature map dimension. As a result, the Kaggle competition's DR dataset has a precision, accuracy, sensitivity, and specificity of 0.974, 0.974, 0.974, and 0.992. According to several experiments, InceptionV4, depending on the ECA, could more accurately detect disease features and define DR severity.

1. Introduction

Diabetic retinopathy (DR) is a serious complication of diabetes caused by capillary rupture triggered by high blood sugar levels [1]. Diabetes affects 460 million people aged 20 to 79 worldwide, and this figure is anticipated to climb to 700 million by 2045 [2], [3]. The number of persons living with DR is anticipated to reach 191 million by 2030 [4]. Initial DR is much less dangerous, does not cause significant visual impairment, and is medically treatable [5]. Early treatment can reduce the probability of vision problems by around 57% [6]. As a result, the most important procedures to safeguard visual acuity are quick inspection and treatment.

2D color fundus images and 3D optical coherence tomography (OCT) images are the most often used examination modalities and the diagnostic basis for ocular disorders [7]. Additional abnormalities can be identified by color pictures, such as microaneurysm, optic disc edema, macular edema and, among others. Deep learning has become more widespread in preprocessing and computer-aided diagnoses in recent times as large data and computer technology have evolved.
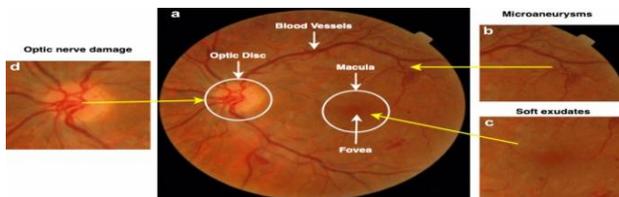

Figure 1: Additional abnormalities can be identified by color pictures.

Use of such fundus pictures to assess the presence of DR related to software assessment can help clinicians make better diagnoses, save money on medical treatment, and offer convenience to poor places. In computer vision, (DCNNs) are broadly used and have been shown to be extremely successful in a wide range of applications. In [8], on the EyePACS-1 and Messidor-2 datasets, the inception-v3-based DCNN has 90.3 percent sensitivity and 98.1 percent specificity for determining DR severity. Authors in [9], used the Kaggle platform to perform transfer learning and hyperparameter finetuning on AlexNet, VggNet, GoogLeNet, and ResNet, with the greatest classification accuracy of 95.68%. Gadekallu et al. are a group of researchers that have come up with a novel way to solve. Previous studies, according to [10], lacked data image preprocessing and reduction in dimensionality, resulting in poor results. They proposed a feature extraction technique that combined principal component analysis, standard scalar data, and DNNs; the model was then particularly different from current mainstream machine-learning models.

Because many intricate elements exist in color fundus pictures like capillaries, which are scattered throughout most of the image, correct preprocessing of the original image is essential when using DCNN for DR classification. More crucially, when employing DCNN for DR classification, the network model must evaluate the significance of the original picture in different locations and channels. This often necessitates the addition of an appropriate attention mechanism, allowing the prototype to dynamically enhance perception [12]. To modify the weights within each function blocks adaptively, the attention method with a compact bilinear pooling for DR fine-grained is utilized. The channel attention mechanism has recently been shown to have a lot of promise in terms of boosting the performance of DCNNs.

Most methodologies, on the other hand, focus on building more advanced attention modules to boost efficiency, which inevitably raises computational complexity. Deep-learning architectures like as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks have been utilized in disciplines such as computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate research, material inspection, and board game programming, generating results that are equivalent to, and in some cases, superior than, traditional approaches. This work proposes an Efficient Channel Attention (ECA) module with only a few parameters and a large performance gain to tackle the performance-complexity trade-off dilemma. By analyzing the channel attention module in SENet, we show that avoiding dimensionality reduction is crucial for learning channel attention, and that appropriate cross-channel interaction may keep performance while substantially decreasing model complexity. As a result, we present a local cross-channel interaction technique that does not require dimensionality reduction and can be performed effectively using 1D convolution. Additionally, we propose a method for estimating coverage of local cross-channel interaction by adjusting the kernel size of 1D convolution. The suggested ECA module is both effective and efficient; for example, when compared to ResNet50's backbone, our modules' parameters and computations are 80 vs. 24.37M and 4.7e-4 GFLOPs vs. 3.86 GFLOPs, respectively, with a performance jump of more than 2% in Top-1 accuracy. We comprehensively evaluate our ECA module on image classification, instance segmentation, and object recognition using the ResNets and MobileNetV2 backbones. According to the results of the trials, our module is more effective and functions better than its competitors.

2. Related work

The goal of DR severity recognition is to assist physicians in making an early diagnosis of early fundus disease and providing a rationale for further diagnosis based on the severity of DR by OCT images using image preprocessing techniques and computer technology or distinguishing lesion features on colour fundus images. To detect lesion traits in the early stages of DR research, researchers mostly relied on typical machine learning approaches. However, as artificial intelligence and computer technology have advanced in recent years, a growth number of researchers have turned to deep learning approaches to classify DR severity.

At the level of DR detection using standard machine learning techniques, researchers must have some medical understanding. Theoretical and manual extracting of lesion aspects from the image collection, followed by feeding the retrieved lesion characteristics into a DR detection. For DR severity classification, the suggested multilayer feedforward NN with high resilience is proposed in [13]. The SVM technique is utilizied to categorize preprocessed bright exudates, non-lesion regions, and cotton wool spots for early categorization and identification of the primary symptoms of DR is suggested in [14]. To identify fundus bleeding, a developed a top-down technique, offered mixed 2DPCA, and

used virtual SVM to improve categorization accuracy is demonstrated in [15].

The authors of [17] explored lesion characteristics using morphological processing, image preprocessing, and texture analysis schemes, that were then fed into an ANN for autonomous DR severity diagnosis. Barriga et al. offer an automated technique for assessing DR lesions [18]. The system employed amplitude and frequency modulation to extract features and used partial least squares (PLS) and a SVM for categorization. In [19], Priya and her associates examined the performance of a probabilistic NN and a SVM to DR figure and found that the SVM outperformed the other models with a 97.608 percent accuracy. Unlike conventional techniques, this article presents a novel attention mechanism to tune the adaptable convolutional kernel sizing method depending on the size of the input feature matrix and fuse the local channel correlation of the feature map to obtain the global channel correlation before combining it with a DCNN for DR classification. This study offers a heatmap to depict the region to which the model pays close attention to evaluate the model's ability to differentiate lesion features at various stages.

3. Proposed Method

3.1. Deep learning

Deep learning automatically extracts the features from the input data using multiple-level of abstraction. Deep learning tries to find the important information in the input images, allowing the classifier to archive high classification accuracy. The number of the proposed network parameters is shown in Table 1.

Table 1: Comparisons with state-of-the-art CNNs on ImageNet

ECA-Net50(ours)	24.37M	3.86G	77.48	93.68
ECA-Net101(ours)	42.49M	7.35G	78.65	94.34

ANNs were inspired by the processing of information and the spread of communication nodes of biological systems. ANNs differ from biological brains in several ways. Artificial neural networks are fixed and symbols, whereas the biological brains of most living creatures are dynamic (plastic) and analog.

3.2. Attention mechanism

Treisman et al. [20] presented the attention mechanism to simulate a human brain attention model that can generate attention weights for many parameters, highlighting the effect of a specific component on the model's outcomes. Deep learning tasks that have used the attention mechanism include sequence [21], picture localization, image comprehension [22], and lip translation [23]. The Google Machine Vision team's proposed transformer architecture [24], which reduces recursion and convolution structures in the best interest of a simpler attention mechanism for accessing feature sequence data, achieved 28.4 BLEU in the WMT 2014 English-to-German translation, which was 2 BLEU higher than the best outcome at the time.

In recent years, the attention mechanism has been progressively coupled with computer vision to adaptively change the higher order abstract information recovered by the model for greater performance. Figure 2 shows the architecture of the attention mechanism.
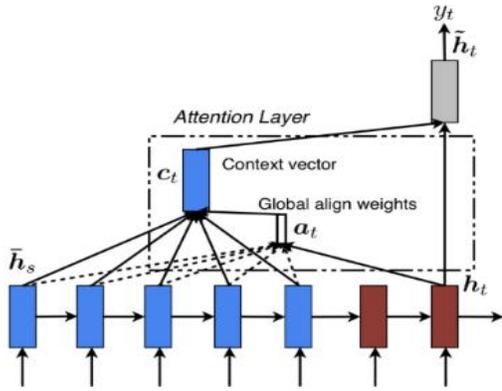

Figure 2: Attention mechanism Structure

3.3. ECA-Net STRUCTURE

Channel attention is one of the robust attention mechanisms that minimizes the model complexity without reducing dimensionality, e.g., the squeeze-and-excite attention mechanism.

This ECA tracks local cross-channel interaction by evaluating each channel and its K partners using channel-wise global average pooling with no dimension reduction. The ECA may be formed using rapid 1D convolution of size K, where K defines the coverage of local cross-channel interaction as well as how many neighbors are involved in attention forecasting for one channel. Figure 3 shows the ECA architecture. Figure 4 shows the proposed framework.
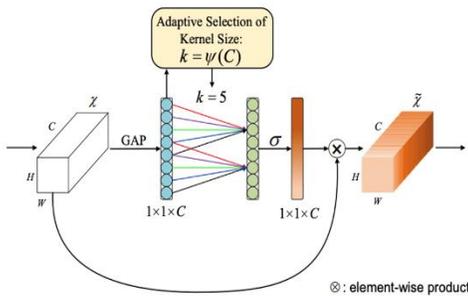

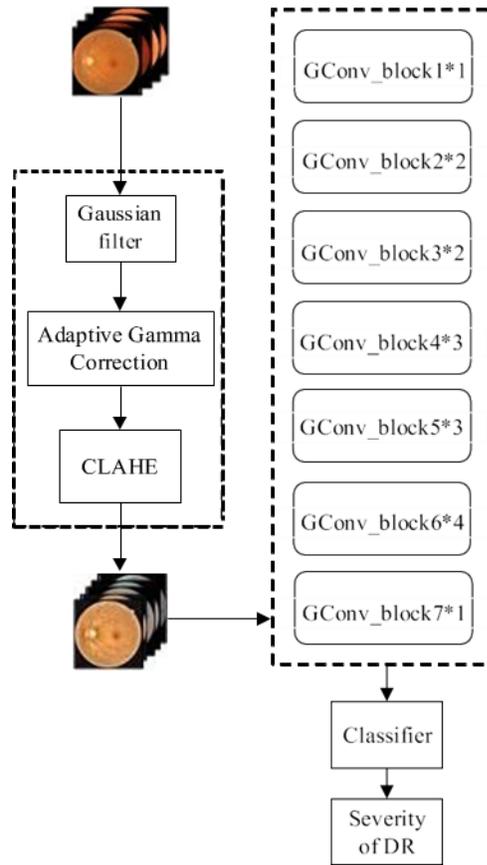*Figure 3: ECA's structure is depicted in this figure.*

input

Figure 4: The architecture of the proposed framework.

output

4. EXPERIMENT

4.1. DATASET

We use the Kaggle challenge dataset to train and test the proposed algorithm. The Kaggle dataset consists of 35126 highly saturated color fundus photos categorized by medical specialists into 5 groups based on the prevalence of DR. Table 2 shows the number of images in each group. The sample size of the various groups in this dataset fluctuates substantially, which will have an adverse effect on the model's findings [25]. The preprocessing strategy will handle this in the subsequent phase.

4.2. AUGMENTATION OF DATA AND PRE-PROCESSING

The Kaggle dataset is collected from retinal photos in different locales, which introduce interference and have negative effects such as irregular illumination, so image pre-processing is required to minimize the impact of noise on the experimental analysis and increase cognitive influence of the network model. Meanwhile, to address the unequal number of DR images in distinct classes, this study employs data augmentation on negative samples in terms of making the samples for each group about equivalent.

Table 2. DR dataset size of the sample

Category	The degree of DR	Number
0	Non-DR	17124
1	Non-Proliferative DR Weak	1485
2	Non-Proliferative Medium DR	3282
3	Non-Proliferative DR High	563
4	Proliferative DR	434

4.2.1. PRE-PROCESSING IMAGES

We perform several preprocessing steps (Figure 4) as follows:

(A) To lessen the impression of extraneous information, remove the black area surrounding the fundus picture.

(B) With Gaussian filtering, it keeps random noise out of the image at capture.

(C) In picture collection, adaptive gamma correction improves contrast and lessens the influence of unequal light. It also fixes images that have less or greyer.

(D) To compensate the impact of unequal grey scale values, intelligent and adaptive balance requires altering the colour space, followed by brightness intelligent and adaptive equalisation. The preprocessed DR pictures are generated by repeating the procedures described above; Figure 5 displays pairs of DR images before and after preprocessing for different categories.

While the pre-processed DR images may be learnt directly for the network, Table 2 shows that the quantity of DR images in each category varies significantly, which will have a detrimental influence on the network model's performance. As a result, this study applies data augmentation strategy, including random rotation, horizontal and vertical rotation, cropping at the four borders and center of the negative sample pictures, and ensuring the required sample size is met. Table 3 compares sample numbers in between data augmentation.

Table 3: Sample size comparison in between data augmentation.

category	The degree of DR		Number	Augmentation Of Data
0	Non-DR		17124	17124
1	Non-Proliferative	DR Weak	1485	17820
2	Non-Proliferative Medium DR		3282	19692
3	Non-Proliferative	DR High	563	16890
4	Proliferative DR		434	16926

Class0 Class1 Class2 Class3 Class4

Figure 5: Pre-processed and unprocessed DR images are compared. The original image is shown in the first row, and the image after pre-processing is shown in the second row.

5. EXPERIMENT ENVIRONMENT

The proposed ECA-Net runs on Python 3.6 with PyTorch 1.7.0. It separates the set of data in an 8:2 ratio between training and validation. We use the cross-entropy as the loss function with training epochs of 100. We optimize the network parameters using a gradient descent algorithm with a learning rate of 0.01. The cosine annealing learning rate modification approach was utilized in this study to make the model eventually converge.

A key difficulty in medical image processing is the absence of appropriate annotated data. While learning DCNN, a limited training dataset is capable of approximating DL approaches also require substantially more training time and data than traditional machine learning systems. To tackle the issues, this research uses deep neural training method in which a pre-trained network with general properties is utilized in another activity, and just a few properties like borders, forms, and texturing may be shared between computer vision projects. Due to the general utilization of transfer learning algorithms, the pre-trained network may be fine-tuned in regression tasks, significantly lowering training time and data, allowing the model to converge as early as possible, and minimizing training error.

The variables of same architecture of ECA are used to guarantee that the model reaches the required performance as rapidly as feasible.

To confirm that the model reaches the desired efficiency as rapidly as feasible, the variables of the same structural system of ECA-Net as well as EfficientNet-B0 are used to define the ECA-Net network using the learning algorithm in the prototype training phase. The ECA components with various structures are designed to use the Kaiming configuration [26], and the compare methods are user-defined with the formal pre-trained consistencies of Python language. This paper evaluates ECA-Net with classic DCNN networks in the same context to test the efficacy of ECA-Net in DR disease detection, with the results discussed in the next section.
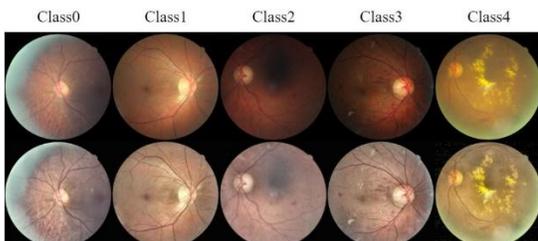
6. EXPERIMENT RESULTS AND ANALYSIS

To thoroughly analyses ECA-DR Net's classifier, the classifier is tested for 50 epochs. The accuracy, sensitivity, precision, and specificity metrics are used to evaluate the performance of the proposed algorithm.

The confusion matrix for DR severity classification is shown in Table 4, and the ECA-Net model is evaluated against the basic DCNN method.

Table 4: Explanation of the Confusion Matrix.

True Value Predict Value	Label			
	Positive Positive	Negative Positive	Positive Negative	Negative Negative
Result	TP	FP	FN	TN

Table 3 shows how accuracy, sensitivity, precision, and specificity are represented:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad , \qquad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP+NP} \qquad , \qquad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad , \qquad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad , \qquad (4)$$

Table 5 shows the classifier for DR disorders of various severity for which ECA-Net obtains an accuracy of 97.4 percent after 100 epochs of training. Figure 6 shows the confusion matrix of the test set.

Table 4: ECA-Net classification performance for DR illnesses of varying severity.

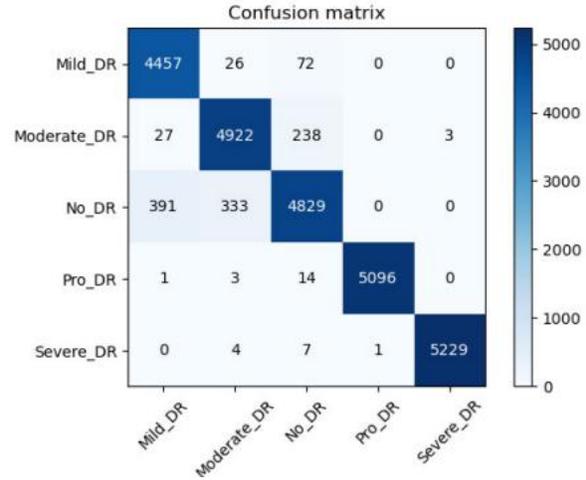| Severity of DR	Sensitivity	Precision	specificity
Non-DR	0.93	0.91	0.96
Non-Proliferative DR Weak	0.91	0.96	0.99
Non-Proliferative Medium DR	0.93	0.94	0.98
Non-Proliferative DR High	0.99	0.99	0.99
Proliferative DR	1.0	0.99	0.99

Figure 6:The confusion matrix for DR disorders of various severity levels the concentration of values on the diagonal is higher, suggesting a good categorization outcome.

Lower-order aspects of the picture, such as curves, borders, and surfaces, are preserved by shallow layer of CNN, while the image's higher-order semantic information is preserved by the deep layer. To have a better understanding of ECA-attention Net's mechanism and how it generates forecasts Internally, multiple depths of blocks are selected to presentation within that work.

Figure 7 shows a heat map of Grad-CAM-based DR images displaying the variable degrees of ECA-Net attention to various sections of the image after using ECA. Red color indicates that image characteristics in the region have greater weights, while blue color suggests that they have lower weights. Figure 7 illustrates this; it can be observed that the feature maps in the model at different levels communicate various details about the diseases seen in the DR images.
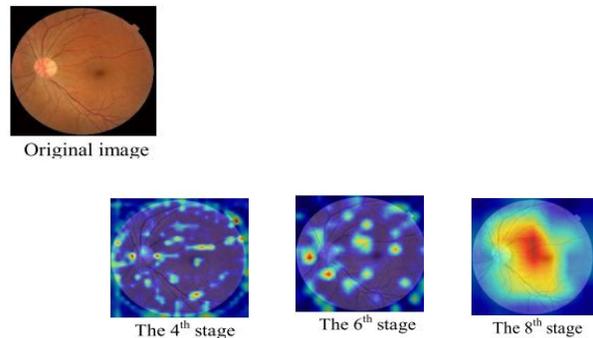

Figure 7: Heat map of the GCA visualization stages 4, 6, and 8 in ECA-Net.

Table 5: ECA-Net results were compared to those of other DCNN models. The more bolded the text, the higher the performance outcomes.

Models	Accuracy	Precision	Sensitivity	specificity
DenseNet-121	0.923	0.926	0.92	0.976
ResNet -101	0.862	0.87	0.88	0.945
GoogleNet	0.91	0.91	0.914	0.965
ECA-Net	**0.974**	**0.974**	**0.974**	**0.992**

In this study, we also perform a comparative test comparing ECA-Net and the traditional DCNN model in the identical experiment setup as described above, analysing accuracy, precision, sensitivity, and specific after 100 training epochs. Table 6 shows the comparsion results. Table 6 shows that all rating indices of the ECA-Net network outperform the example of the importance CNN network, showing that the DCNN system that is based on ECA-efficient Attention presented in this study may outperform the standard CNN network in the illness degree categorization job.

6.1. RESULTS

Diabetic retinopathy is among the most significant complications of diabetes mellitus, and if not detected and treated early, it can cause loss of vision or perhaps even blindness. Diabetic retinopathy, on the other hand, may be avoided with frequent screening and appropriate treatment, averting irreparable blindness. As computer vision and artificial intelligence techniques progress, In the health care industry, an increasing variety of algorithms are being employed to aid physicians with everyday therapy. Therefore, the ECA is presented as for feature maps in this paper. In comparison, a DCNN model called ECA-Net For such rapid recognition of diabetic retinopathy, a combination of the ECA and EfficientNet is proposed.

7. Conclusions

This paper proposes a methodology convolutional kernel size adjustment methodology for trying to extract network channel correlation during the disease feature extraction stage that also allows ECA-Net to flexibly modify the convolutional kernel size in various tasks, allowing the model to carefully evaluate the similarity between feature map channels and improve the results.

Transfer learning and parametric annealing methods are employed inside the training process to ensure the model converges as quickly as possible. On the DR validation set, the final ECA-Net model obtains accuracy, precision, sensitivity, and specificity of 0.974, 0.974, 0.974, and 0.992, respectively. These results demonstrate that the DCNN model according to the ECA attention mechanism presented in this research is likely to succeed. In terms of DR concerns, we hope to combine the ECA with more deep learning models in the future to increase the model's ability to recognize minor variations between classes and enable ECA-Net to be utilized in a different environment.

8. REFERENCES

[1] C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, ''Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,'' Ophthalmology, vol. 110, no. 9, pp. 1677–1682, May 2019.

[2] P. Saeedi, P. Salpea, S. Karuranga, I. Petersohn, B. Malanda, E. W. Gregg, N. Unwin, S. H. Wild, and R. Williams, ''Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the international diabetes federation diabetes atlas, 9th edition,'' Diabetes Res, Clin. Pract., vol. 162, pp.180-186, Apr. 2020.

[3] C. Sabanayagam, R. Banu, M. L. Chee, R. Lee, Y. X. Wang, G. Tan, J. B. Jonas, L. Lamoureux, C.-Y. Cheng, B. E. K. Klein, P. Mitchell, and R. Klein, ''Incidence and progression of diabetic retinopathy: A systematic review,'' Lancet Diabetes Endocrinol., vol. 7, pp. 140–149, 2019.

[4] D. S. Ting, G. C. Cheung, and T. Y. Wong, ''Diabetic retinopathy: Global prevalence, major risk factors, screening practices and public health challenges: A review,'' Clin. Exp. Ophthalmol., vol. 44, no. 4, pp. 260–277, May 2016.

[5] N. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, ''IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,'' Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018.

[6] Early Treatment Diabetic Retinopathy Study Research Group, ''Grading diabetic retinopathy from stereoscopic color fundus photographs— An extension of the modified airlie house classification: ETDRS report number 10,'' Ophthalmology, vol. 127, pp. S99–S119, Apr. 2020.

[7] M. D. Abràmoff, M. K. Garvin, and M. Sonka, ''Retinal imaging and image analysis,'' IEEE Rev. Biomed. Eng., vol. 3, pp. 169–208, 2010.

[8] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, and S. Venugopalan, ''Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,'' JAMA, vol. 316, pp. 2402–2410, 2016.

[9] S. Wan, Y. Liang, and Y. Zhang, ''Deep convolutional neural networks for diabetic retinopathy detection by image classification,'' Comput. Electr. Eng., vol. 72, pp. 274–282, Nov. 2018.

[10] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, ''Early detection of diabetic retinopathy using PCA-firefly based deep learning model,'' Electronics, vol. 9, p. 274, Feb. 2020.

[11] P. Liu, X. Yang, B. Jin, and Q. Zhou, ''Diabetic retinal grading using attention-based bilinear convolutional neural network and complement cross entropy,'' Entropy, vol. 23, pp. 816, Jun. 2021.

[12] R. P. R. Priya and P. Aruna, ''SVM and neural network-based diagnosis of diabetic retinopathy,'' Int. J. Comput. Appl., vol. 41, pp. 6–12, Mar. 2012.

[13] A. Treisman, ''Features and objects in visual processing,'' Sci. Amer., vol. 255, pp. 114–125, Nov. 1986.

[14] T. Bluche, ''Joint line segmentation and transcription for end-to-end hand-written paragraph recognition,'' in Proc. NIPS, vol.215, pp. 838–846,2018.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, ''Spatial transformer networks,'' presented at the 28th Int. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, vol. 9, pp.155-162, 2015.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' vol.131, pp.274-280, 2015.

[17] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ''Lip reading sentences in the wild,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol.220, pp. 3444–3453, Jul 2018.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' presented at the 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, vol.181, pp.298-315, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, ''Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,'' vol.19, pp.150-155,2015.

[20] A. Treisman, ''Features and objects in visual processing,'' Sci. Amer., vol. 255, pp. 114–125, Nov. 1986.

[21] T. Bluche, ''Joint line segmentation and transcription for end-to-end handwritten paragraph recognition,'' in Proc. NIPS, vol.19, pp. 838–846,2016.

[22] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, ''Spatial transformer networks,'' presented at the 28th Int. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, 2015.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' 2015, arXiv:1502.03044.

[24] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ''Lip reading sentences in the wild,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3444–3453.

[25] https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' presented at the 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.