IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 6, December, 2023, pp. 1617-1627
Journal homepage: http://iieta.org/journals/ria

Image Transformers for Diabetic Retinopathy Detection from Fundus Datasets

Nikhil Sathya Kumar[1*], Ramaswamy Karthikeyan Balasubramanian[2], Manoj Ravindra Phirke[1]

[1] Imaging and Robotics Lab, HCL Technologies Ltd, Bangalore 562106, India
[2] Department of Electronics and Communication Engineering, M. S. Ramaiah University of Applied Sciences, Bangalore 560058, India

Corresponding Author Email: ss.nikhil.sss@gmail.com

https://doi.org/10.18280/ria.370626

Received: 17 August 2023
Revised: 1 October 2023
Accepted: 8 October 2023
Available online: 27 December 2023

Keywords:
diabetic retinopathy, image transformers, convolutional neural networks, mixer architecture, APTOS dataset, F1-score, age related macular degeneration

ABSTRACT

Diabetic retinopathy (DR) is the major cause for blindness worldwide. Early DR detection is crucial for preventing severe vision loss. Timely interventions improves patient outcomes, hence clinicians advice diabetic patients to undergo periodic retinal screening using fundus cameras, where DR can be identified through distinct retinal biomarkers like hemorrhages, aneurysms and exudates. These biomarkers can be detected using Deep learning techniques like Convolutional-Neural-Networks (CNN), Vision Transformers and Mixer architectures. The objective of this paper, is to develop, investigate and identify the architectures and algorithms that relatively improves the detection of DR. In this paper, 18 pre-trained state-of-the-art opensource models like ResNet, EfficientNet, BeiT, VOLO, TNT, DeiT, Visformer, CoAT-NET, CaiT, XCiT, Poolformer, Swin, Twin, PiT, MLP-Mixer, ResMLP, ConvMixer were used for DR detection. A custom classification-head containing Global Average Pooling layer, Fully Connected layers, Dropout, Activation functions and Softmax layer were added to pre-trained models. The entire architecture was fine-tuned, evaluated and benchmarked on multiple opensource fundus datasets using NVIDIA-GeForce-GTX-1080 GPU. Different hyperparameters like batch-sizes, normalization, dropout, activation, optimizers and learning-rate-decay functions were evaluated to improve the performance of the models. Overall, around 71 different experiments were conducted to achieve state-of-the-art F1-scores of 99.3%, 88.7%, 85.25%, 64.16%, 86.52% and 90.53% for APTOS-DR detection, APTOS-DR grading, Messidor, IDRiD-DR, IDRiD-AMD and AREDS datasets respectively, which was around 2% better than current state-of-the-art. Performance of Transformers was better than CNN and Mixer based architectures because of their ability to learn the global context and associate position of biomarkers with other anatomies of retina. F1-scores of Swin, PiT and Twin models were highest among all the Transformers because of their ability to encode fine as well as coarse-level details of biomarkers.

1. INTRODUCTION

More than 40 million individuals in the US alone suffer from serious eye related ailments that, if left untreated, can result in total blindness. A large number of these eye diseases affect the retina. Diabetic Retinopathy (DR), Age-Related Macular Degeneration (AMD) and Glaucoma are the most common retinal diseases. By 2014, diabetes had affected more than 422 million people. Nearly all individuals with type-1 diabetes and 60% of those with type-2 diabetes are predicted to contact diabetic retinopathy within the first 20 years of having the disease.

Ophthalmologists frequently suggest diabetic individuals to undergo periodic medical screening for an early diagnosis of DR. Lot of studies have exhibited that DR can be treated effectively if it is detected early. Conventional approaches to diagnose DR, required manual grading of the retina to detect the absence or presence of the disease. This kind of diagnosis is time consuming and expensive, as it requires human expertise. Automated disease detection makes disease diagnosis more accessible to a larger population. The ability of computer-aided systems to consistently and accurately grade DR has popularized them among researchers.

Fundus cameras are the most popular devices for large scale population screening of retinal images because they are non-invasive, accurate, consistent, easy to operate, cost effective, provides storage options for the images, and contain established biomarkers for DR detection. Major biomarkers for the detection of DR like Hemorrhages, Micro-Aneurysms, Hard Exudates and Soft Exudates as shown in the Figure 1, can be easily identified with fundus imaging. Clinicians have graded DR into 5 different stages based on the severity namely: Proliferative, Severe, Moderate, Mild and No DR.

Traditionally, computer vision techniques like Haar, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT) were used for feature extraction and Support Vector Machine (SVM), Random Forest, AdaBoost based algorithms were commonly used as classifiers for detection of DR. But the disadvantages of computer vision techniques were lower accuracy and requirement of handcrafted feature extractors, thereby leading to lower adaptability of the algorithms on different kinds of fundus images.

Deep learning based approaches have been proposed to overcome these drawbacks of computer vision techniques. Few of the main advantages of deep learning based architectures are: 1) Automatic feature extractors without the requirement for hand-crafting, 2) Feature extractor and classifier are bundled into a single workflow, 3) Higher accuracy, 4) Model robustness to adapt to different fundus datasets. There are various deep learning architectures like CNN, Transformers and Mixer models that have been proposed in literature. Transformers and Mixer based models are the current state-of-the-art for ImageNet based image classification tasks, but they have not been evaluated and benchmarked extensively on fundus images for retinal disease detection. CNN based models are widely used for DR detection and image classification.

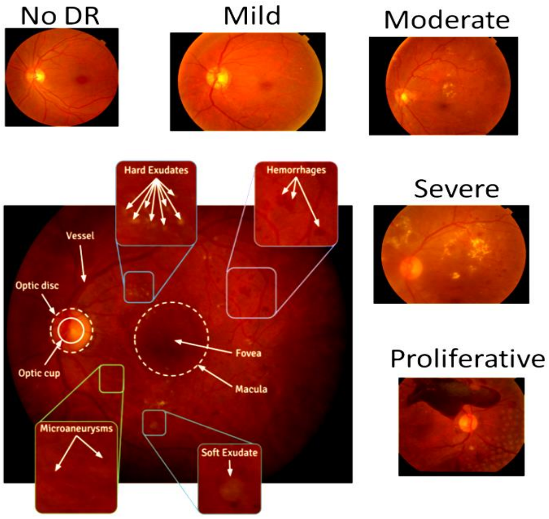

Figure 1. DR biomarkers and grades

CNNs are an extension of neural networks which have been optimized for images. The input images in a CNN model are usually streamed through multiple convolutional layers, activation layers, pooling layers, flattening and fully connected layers to obtain class probabilities. AlexNet [1], in 2012 was the first CNN based model to outperform computer vision-based techniques in the ImageNet image classification competition. Subsequently, CNN-based models continued to win the ImageNet challenge until 2020. Since 2021, Transformer-based architectures exhibit superior performance compared to CNN models in the ImageNet challenge. Image transformers were first proposed by authors of ViT [2] in 2020, and they were inspired from NLP based transformers. In Vision Transformers, the input image is divided into patches and converted to a 1d array by a dense layer. Positional embeddings are added to each of the patches, and streamed through multiple transformer encoders which contain multiple normalization layers, Multi-head attention layers and dense fully connected layers. Finally the classification scores are derived using a final fully connected layer. Mixer based architectures [3] were first proposed in 2021, they suggest to replace convolutional layers in CNN and attention networks in Transformer with MLPs. Each of these architectures have varied features that are beneficial for different classification problems.

In this paper, multiple pre-trained models based on CNN, Transformer and Mixer architectures were chosen for analysis. The models were chosen based on accuracy when trained on ImageNet dataset, and in order to allow for comparable depiction of various architectures. CNN architecture based ResNet and EfficientNet were chosen for evaluation. MLP-Mixer, ResMLP and ConvMixer based on Mixer architectures were also evaluated, along with Vision Transformer (ViT) and its variants. The modification to the Transformer architecture included: 1) Dividing the patches in ViT model to sub-patches for extracting the finer features from the input image, 2) Attention network related changes in the ViT model, 3) Modification of Transformer architecture to include CNN based features, 4) Auto-encoder related design for Transformer network. Models containing a fusion of these modifications, like Pooling based Vision Transformer (PiT), Swin and Twin Transformer achieved the best performance when compared to all other architectures.

The pre-trained models were coupled with a custom classification-head, which contained 2 fully connected layers, ReLU based activation function, AdamW optimizer, softmax activation and exponential multi-step learning rate decay function. Complete model was fine-tuned with APTOS dataset.

The Key Performance Indicators (KPIs) used for benchmarking the models were: 1) F1-score to indicate the DR detection capability of the model, 2) Number of parameters of the model, and 3) Time required to train and infer from the models. APTOS dataset was chosen for extensive ablation study.

The primary role of transformers architectures are attention mechanism and position encoding. The attention module and the position encoding captures global information of the image, thereby yielding global interaction between image patches and enables flexible modeling of image data beyond local interactions of convolutions (in CNN and Mixer models). Hence the F1-score for Transformer based models was higher than CNN and Mixer models. Transformer design requires fewer convolutional operations than CNN and Mixer models, which results in a smaller model size and shorter training and inference times. The Transformer based models have the capability to improve their performance when trained on larger datasets. This paper exhibits the comparative study of Transformers, CNN and Mixer Architectures for DR detection and their relative performance analysis.

In the initial ablation study, 34 different experiments were conducted like: a) 3 different number of fully connected layers, b) 3 batch sizes, c) 7 variants of ReLU activation function, d) 5 variants of Adam optimizer, e) 10 learning rate decay functions, f) 3 dropout function g) 2 normalization functions.

2 fully connected layers, batch size of 32, ImageNet based normalization, dropout function, ReLU based activation, RAdam optimizer and “Multi-step Learning Rate (LR) decay” were found to be the most suitable hyperparameters for DR detection. With these settings, the selected models and techniques were tested on other opensource retinal datasets like Messidor, AREDS and IDRiD. The Transformer models resulted in F1-scores which were all better than results reported in the literature reviewed in this paper.

The following sections of this paper is structured as follows: Section 2, provides a summary of the latest literature on retinal disease detection. Section 3, describes the methodologies used including a note on datasets, models and the architecture proposed in this paper. Section 4 contains comprehensive experimental analysis. Finally Section 5 summarizes the paper's content and provides conclusion and a discussion on future work.

2. RELATED WORK

Researchers favor computer-aided detection systems because of their ability to precisely identify the grades of DR. Over the past ten years, many studies have been undertaken to build computer-aided systems that can automatically diagnose DR using machine learning approaches. This section discusses the most recent research on retinal disease detection using different fundus datasets. A summary of the performance of various state-of-the-art techniques for retinal disease detection has been shown in Figure 2.

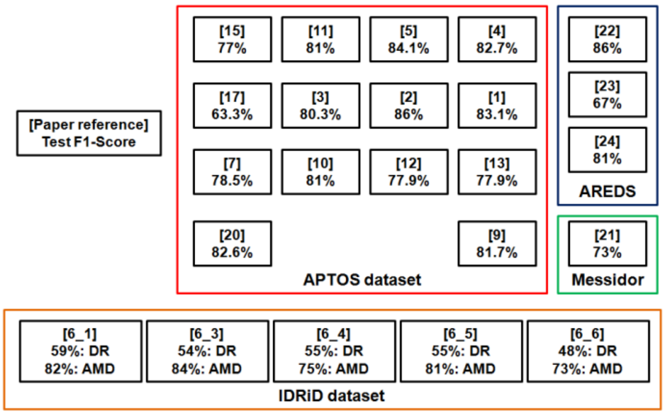

Figure 2. Performance of state-of-the-art for DR detection

Few of the authors propose to use custom CNN based architectures for detection of DR and AMD. Pasha et al. [4], Saleh et al. [5] and Dekhil et al. [6] suggest a CNN model which consists of a pre-processing stage, five stage custom convolutional layers, ReLU based activation followed by pooling layers and three fully connected layers. Liu et al. [7] developed a Graph Convolutional Network (GCN), Global average pooling (GAP) layer and fully connected layers. Zhuang et al. [8] propose a custom 9 layer architecture for processing input images of size 512x512. The custom layers include multiple 2d convolutional layers, activation layers, max-pooling and fully connected layers. Alyoubi et al. [9] suggests using a cascade of 3 CNN architectures containing 2 custom CNNs (for 2 input resolutions) and 1 EfficientNetB0 model. Porwal et al. [10] suggest a custom CNN network, which used the EyePacs dataset for pre-training and IDRiD dataset for fine-tuning, thereby increasing the learn-ability of the model. Seoud et al. [11] suggest the development of handcrafted feature extraction techniques based on the location, size and texture of the biomarkers and classified using random forest. González-Gonzalo et al. [12] propose to stream input images with and without preprocessing to 6-custom CNN models separately. The performance of custom CNN models is relatively lower compared to other architectures.

There are many papers based on the usage of pre-trained networks like VGG-19, InceptionNet and ResNet. Singh et al. [13] propose to use a pre-trained VGG-19 with L-BFGS rather than ADAM optimizer. Pre-trained InceptionNetV3 with momentum based optimizer was suggest by Wang et al. [14]. Peng et al. [15] propose DeepSeeNet, which consists of 3 Inception-v3 models to detect drusen (biomarker for AMD) in 3 size categories, absence or presence of abnormalities and late AMD. Liu et al. [7] recommend 18-layer ResNet. Dondeti et al. [16] and Nasir et al. [17] suggest ResNet based model with deep layer aggregation.

Few of the authors suggest to use DenseNet, XceptionNet, and MobileNet, which are extensions of VGG-19 and ResNet. Chaturvedi et al. [18] propose pre-trained DenseNet121, GlobalAveragePooling2D, dropout and Softmax. A pre-trained DenseNet121 was used with augmentations like translation, rotation, horizontal flip, vertical flip, scaling, Gaussian noise, random blurring and shear by Sheikh et al. [19]. Porwal et al. [10] recommend a pre-trained DenseNet for AMD and DR grading. Kassani et al. [20] suggest a modified Xception architecture along with a deep layer aggregation to combine multi-level features from different CNN layers of Xception network. Wang et al. [21] and Patel et al. [22] suggest MobileNet architecture with GAP layer, fully connected layer and Softmax activations.

EfficientNet, which is considered as an extension of models like ResNet, DenseNet, InceptionNet has been proposed by multiple authors. Dondeti et al. [16] suggest EfficientNet, v-SVM (Support Vector Machine) and t-SNE to avoid over-fitting of models. Zhuang et al. [8] suggested a pretrained Efficientnet-B3 along with a 2D adaptive average pooling layer, dropout layer and a linear layer. Pour et al. [23], propose to stream equalization based pre-processed images through an Efficientnet-B5 model. Xie et al. [24] propose, three individual pre-trained EfficientNetB0, global average pooling layer, dropout, fully connected layer, softmax to detect each of AMD stage, drusen size and presence of pigmentary abnormalities.

There are lot of papers, where the focus is on using cascades of state-of-the-art models. Bodapati et al. [25] suggest a composite deep neural network with gated-attention mechanism, where XceptionNet and VGG16 model are blended using multi-modal fusion. Fusion of AlexNet and GoogLeNet was proposed by Porwal et al. [10]. They also propose an ensemble containing pre-trained ResNets and DenseNets for the grading of DR.

Image transformers have recently attracted a growing amount of interest in computer vision and medical image analysis, producing state-of-the-art results on a number of image classification applications. Mutava et al. [26], Wu et al. [27] and Matsoukas et al. [28] analyze ViT architectures for DR grading, and show that they are competitive alternatives to CNNs. Yao et al. [29, 30] propose Swin based transformers, which is a variant of ViT. Jin et al. [31] proposes a dual-path reasoning network based on transformers which can infer from appearance and geometric features based on the clues discovered by the detector.

This paper intends to address the specific gaps in the existing literature by: 1) Training, evaluating and benchmarking 18 models based on CNN, Transformer and Mixer architectures to detect DR and AMD on multiple opensource datasets, 2) Determining the optimum hyperparameters like number of layers, batch-sizes, normalization factor, activation function, optimizers and learning rate schedulers.

3. RESEARCH METHODS

This section contains a detailed description of: 1) Different deep learning model architectures which have been used in this paper, which includes multiple CNN, Transformer and Mixer models, 2) Retinal fundus datasets that have been considered for model evaluation and benchmarking, 3) Architecture of the overall model that has been proposed in this paper for retinal disease detection.

3.1 Deep learning architectures (CNN vs. Transformers vs. Mixer)

18 pre-trained models have been used in this paper for evaluation and benchmarking and determining the most appropriate model for the detection of DR. The models were selected based on the accuracy on ImageNet dataset and the model structure, so that different architectures have similar number of representations for benchmarking.

CNN based models have translation equivariance. For example, an object can be rotated in the image, but CNNs can accurately detect the object. Leveraging translation equivariance and augmentation, CNNs can be trained with relatively less images compared to other architectures. ResNet proposed by He et al. [32], allows adding more number of layers to neural network through skip connections, thereby avoiding exploding and vanishing gradients and improving the accuracy of the model. In Efficient-Net, Tan et al. [33] propose progressive training for adaptively adjusting the regularization and Neural Architecture Search (NAS) to drastically reduce the time required for training the model and to improve the accuracy.

ViT, vision transformers proposed in the study [2] was the first Transformer model, used for vision applications. Here the input image is divided into several patches and streamed through a dense linear layer, along with positional embeddings and class embedding (CLS Token). These are streamed through multiple normalization layers, attention layers, and dense layers which results in classification scores.

BEiT (Bidirectional Encoder representation from Image Transformers) proposed in the study [34] introduced autoencoder based techniques to improve the performance of transformers and achieve the highest accuracy on ImageNet dataset.

In ViT, the input images are divided into several local patches and their relationships and representations are calculated. But images have abundant color information and are highly complex. The original patch is not granular enough to extract adequate features of objects in different locations and scales. Hence, architectures like Vision Outlooker (VOLO) [35] and Transformer in Transformer (TNT) [36] have been proposed, where the patches are divided into smaller patches and these sub-patches are streamed to a stack of sub-transformers for further processing. This results in efficient encoding of finer-level features and contexts into tokens.

Few architectures like Data Efficient Image Transformer (DeiT) [37], Visformer [38] and CoATNET [39] use Vision Transformers as the base architecture and modify it based on CNN techniques. Distillation networks have been used in DeiT, attention layers have been replaced by convolutional layers and pooling layers by Visformer and CoATNET respectively. By fusing Vision Transformers and CNN, model capacity and generalization is proposed to improve.

Self attention module in ViT leads to global interaction between tokens (image patches), but disadvantage is the quadratic complexity in time and memory. Hence it is not scalable to high resolution images. Authors propose to modify the attention module of the transformers to improve the performance. Few of the models based on this alteration are: XCiT [40], CaiT (Class attention in image transformers) [41] and PoolFormer [42]. In XciT, attention has been proposed over the channels, hence the model has faster inference for high resolution images. Apart from self attention in ViT, class attention was also proposed in CaiT. In PoolFormer, attention block in DeiT is replaced with normalization, pooling and MLP layers.

Few transformer based architectures use a fusion of various features on the base ViT model like: 1) Dividing the image patches to sub-patches for obtaining finer as well as coarse-level details of the images, 2) Using CNN based improvisations on the transformer module and 3) Altering the attention layer to achieve better performance. Few of these models are Swin, PiT (Pooling based Vision Transformer) and Twin transformer. In Swin and Twin transformers, the number of channels is increased while the resolution of the input image is decreased similar to CNN models. Attention has been calculated for a patch with respect to all other patches in the entire image as well as a smaller window. This leads to extraction of finer as well as coarse-level features. Twin transformer introduces Spatially Separable Self Attention, as an optimization technique for Swin transformer. PiT model downsizes the input image and increases the number of channels. It uses pooling layer to reduce the spatial size of the input image, thereby improving the generalization and expressiveness of the model.

Mixer based architectures use the idea of convolutions with small kernels. It reduces the kernel size to 1*1, thereby turning convolutions into standard dense matrix multiplications. These are applied independently to each spatial location (channel-mixing MLPs) and to each feature (token-mixing MLPs). The different Mixer architectures that have been used in this paper are MLP-Mixer [3], Res-MLP [43] and ConvMixer [44]. In MLP-Mixer architecture, token mixing and channel mixing has been proposed without CNNs or attention mechanisms. Only MLPs based on matrix multiplications, non-linearities, normalization, skip connections are proposed. Res-MLP is similar to MLP-Mixer with the addition of an affine layer. ConvMixers is a hybrid between ViT and MLP-Mixer.

3.2 Dataset

Among the various opensource fundus datasets that are available for DR detection, APTOS dataset was used for model benchmarking and extensive ablation study for the following reasons: 1) Dataset was released as part of an open competition, hence there are benchmarks available in the public leader-board, for researchers to compare their algorithms, 2) Volume of dataset, 3) Numerous publications that use the APTOS dataset for benchmarking. The images are graded for 5 classes to detect DR. The best performing models on APTOS dataset, were shortlisted and evaluated for robustness on other retinal datasets like Messidor, IDRiD-DR, IDRiD-AMD and AREDS for DR and Age-Related Macular Degeneration (AMD) detection. IDRiD-DR and IDRiD-AMD datasets also have public leader-board, hence it is convenient for researchers to benchmark their results. Messidor and AREDS are other popular opensource retinal datasets, which have been used for model evaluation. The dataset distribution is shown in the Table 1. In order to avoid model overfit, images in the dataset were shuffled and distributed into training set, validation set and test set.

Table 1. Dataset distribution

Dataset	APTOS: DR Unbalance / Balance	Messidor DR	IDRiD DR/AMD	AREDS AMD
Image Size	1050*1050	1440*960	4288*2848	2240*1488
No Disease	1800/1800	546	168/221	974
Mild	368/1545	153	25 / 51	75
Moderate	988/1779	247	168 / 0	0
Severe	191/1425	946	93 / 243	151
Proliferative	293/1231	0	62 / 0	0
Total Images	3640/7780	1892	516/515	1200

3.3 DR detection model architecture

For benchmarking, 18 pre-trained models were selected in this paper based on the model structure and F1-score on ImageNet dataset. Final classification layer of the pre-trained models were removed, hence the pre-trained models were used as feature extractors and they were coupled to a base classification-head. The extracted features are streamed through a classification-head, which predicts the grade of DR. The base classification-head consists of 2 fully connected layers with dropout, ReLU based activation, AdamW optimizer, multi-step learning rate decay function, batch-size of 32 and ImageNet based normalization function. The architecture of the proposed network is shown in Figure 3.

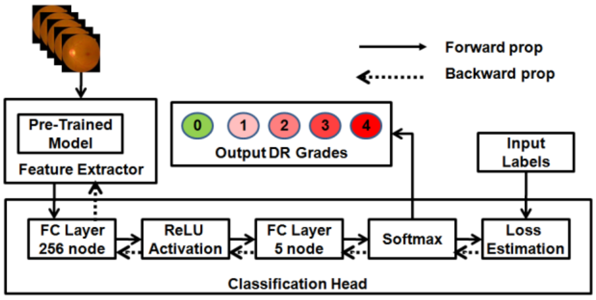

Figure 3. Architecture of proposed DR detection network

The input image resolution for all the models was 224*224, to maintain consistency in benchmarking. The dimensions of the extracted features were reduced to 256-d when they were passed through the first fully connected layer in the classification head. To introduce non-linearity, ReLU based activation layer was used. After the activation layer, another fully connected layer that reduces the dimension from 256-d to 5-d (corresponds to the number of classes) was used. To derive the probability of the classes, softmax layer was used. Dropout with a probability of 30% was added to the fully connected layers to improve the model robustness and reduce the model overfit. Using dropout, certain neurons in the fully connected layer are randomly blocked, thereby reducing the reliance on certain activations. Hence the model will be able to learn more significant features.

At the start of the training process, the weights of the classification-head were initialized randomly and used to convolve with the activations. During forward propagation, a batch of input images were streamed through the different layers, to eventually result in the probabilities of DR grade for each of the input images. After the output probability values were determined from the softmax layer, it was compared with the ground-truth labels and error was estimated using log likelihood loss function. The primary objective of the model training process was to minimize this error, hence the weights of the fully connected layers were updated. During back-propagation, the error was estimated and the weights were updated.

The process of forward and backward propagation was executed over multiple iterations for different batches of input images, till there was a satisfactory decrease in the error. Over different iterations, the weights were updated using AdamW optimizer, and multi-step learning rate decay function. Input images were streamed in batches of 32. To maintain consistency during the benchmarking of 18 pre-trained models, the classification-head and the dataset distribution were used without any modifications.

4. RESULTS AND DISCUSSIONS

This section contains: 1) Detailed analysis of the results when the 18 different models based on CNN, Transformer and Mixer architectures were fine-tuned on APTOS dataset, 2) Extensive ablation study on the 3 models with highest F1-scores to select the optimum methods and hyper-parameters including batch size, normalization function, activation function, optimizer and learning rate decay function, 3) Comprehensive performance analysis of the selected models and methods on other opensource retinal datasets.

4.1 Performance analysis of deep learning architectures on APTOS dataset

Table 2 shows the performance of 18 state-of-the-art pre-trained models when trained on APTOS dataset. The parameters that were considered for evaluating the model performance were F1-score, size of the model, time required for training and inferring from the model. The models were trained on a NVIDIA GeForce GTX 1080 GPU, 64 bit Ubuntu 18.04.6 LTS, 31.2 GiB internal memory and Intel® Core™ i9-8950HK CPU @ 2.90GHz × 12 processor.

The F1-score of CNN based ResNet and EfficientNet models was lower compared to Transformer and Mixer based architectures as they do not capture global information and have lower capacity to learn from larger datasets. Inference time required for CNN based models was higher due to the number of convolutional computations. The training time per epoch for ResNet was the highest compared to all other state-of-the-art models. Architectural improvements made by EfficientNet, led to faster model training time compared to ResNet.

The training time required for Transformers is lower than CNN and Mixer based architectures. In biomedical applications, the number of training images will increase with time, hence model robustness is a key performance indicator. Compared to CNN and Mixer based models, Transformer have higher learning capacity when trained on larger datasets. Main building blocks of transformers are attention module and position information. Attention module and position encoding captures global information of the image, thereby enabling flexible modeling of image data beyond local interactions of convolutions (in CNN and Mixer models). The disadvantage of attention module is the quadratic complexity in time and memory, hence it is not scalable to high resolution images. The disadvantage of global receptive field is that, even if the image is rotated or flipped, it does not cause any issue to the transformer. Hence transformers tend to overfit the data very easily and some augmentation techniques might not actually improve the accuracy. There are various alternatives to ViT, which optimize the model for better performance.

BEiT, is an autoencoder based transformer architecture. Pre-trained BEiT model achieved the highest accuracy on ImageNet dataset, but the model under-performed when fine-tuned on APTOS dataset. The model initially learns to reconstruct the input image, and then to classify it to different classes. During reconstruction, model might add noises to the reconstructed image which can be misinterpreted as biomarkers by the classifier, leading to lower F1-score. Performance of BEiT model would be better, if the classes are distinct (like ImageNet dataset on which it is pre-trained and benchmarked). But if the classes are not distinct enough, then the F1-score would be low.

One of the main disadvantages of ViT architecture was the low efficacy in encoding fine-level features into token representations. Modifications of ViT transformers like TNT and VOLO, where the patches were divided into sub-patches (similar to how sentences are divided into words) have been evaluated. This enables the transformers to learn finer details from the fundus images, thereby improving the F1-score of TNT and VOLO compared to ViT. In VOLO, as the patches are divided into sub-patches, and these sub-patches are used for training, hence the training time is higher compared to ViT model. Inner transformer architecture proposed by TNT model, performs better than VOLO in terms of lower model size (as the parameters within the inner transformer are shared), higher F1-score and lower training time. F1-score of VOLO and TNT is lower compared to other transformers based architectures.

Fusion of Transformer and CNN based techniques achieve better performance. Few models that belong to this category are DeiT, Visformer and CoATNET. The addition of distillation token, has improved the F1-score of DeiT model. The training time required for this model is also comparable to few of the fastest transformer architectures. The model size is also significantly smaller than ViT. Visformer is a transition from DeiT towards ResNet based model, where the positional embeddings are removed from DeiT model and replaced with feed forward network. This change seems to have degraded the F1-score, thereby signifying the importance of positional embeddings for DR detection. Visformer has the fastest training times compared to all other state-of-the-art models. CoATNET is a combination of transformer and CNN models stacked together. Because of multiple training heads, the overall training time was higher compared to other transformer based models. The resolution of the input image was progressively reduced throughout the model, but the attention module was computed only on the patches, rather than the sub-patches. This led to loss of information especially with regards to biomarkers, thereby leading to slightly lower F1-score compared to best performing model for the detection of DR.

Table 2. Results of 18 state-of-the-art models when trained on APTOS dataset

Batch 32, ImageNet norm, ReLU, AdamW, multi-step LR decay, 20 epochs, cross entropy loss, LR 3e-4					
Model Category	**Model**	**Model Size (MB)**	**Test F1-Score (%)**	**Train Per Epoch (min)**	**Infer Time ms**
CNN based architectures					
CNN	ResNet	468	83.3	41.1	124
	Efficient-Net	9.11	81.1	5.8	88
Transformer based architectures					
Original Transformer	ViT	304.3	83.9	12.4	42
Transformer + Autoencoder	BEiT	304.4	80.3	12.8	53
Patches into sub patches	VOLO	295.4	84.8	20.4	23
	TNT	23.7	85.3	4.9	22
Transformer + CNN	DeiT	87.34	86.2	3.9	23
	Visformer	40.2	85.9	2.5	21
	CoAT-NET	10.3	86.1	8.4	22
Transformer, Attention changes	CaiT	46.9	85.3	4.4	24
	XCiT	188.9	85.4	39.8	12
	Pool-Former	73.4	86.2	4.6	23
Transformer, sub-patch, CNN, Attention	Swin	196.5	88.1	8	53
	Twin	43.8	87.0	3.4	21
	PiT	74.7	86.9	3.6	22
Mixer based architectures					
Mixer	MLP-Mixer	59.8	81.5	3.2	27
	ResMLP	129.1	84.0	18.3	74
	Conv-Mixer	51.6	80.7	20.3	63

Changes to attention network in transformer based architectures were proposed by CaiT, XCiT and PoolFormer models. In CaiT model, 2 different stages have been proposed, namely: class attention and self attention. During self attention (wherein, attention within sub patches was calculated), class embeddings were not considered. Hence the F1-score of CaiT model was lower than DeiT, though these class embeddings were reintroduced during class attention. In XCiT, instead of attention over image patches, attention is proposed over the channels. This modification has made the network faster in terms of inference time, though it takes longer for training. This network is ideal for high resolution images. Though the

F1-score is comparable to most of the transformer models, it is slightly lower than the best performing models. PoolFormer is a modification of DeiT, where attention module is replaced by a pooling layer. Most of the parameters like F1-score of the model, training time, inference time are almost similar between DeiT and PoolFormer. This shows that there is very little difference between attention module and PoolFormer layer. The number of parameters in PoolFormer has been reduced compared to DeiT, because attention block was replaced by pooling and the pooling layer does not require any learnable parameters.

Swin, Twin and PiT transformer are Transformer based architectures which are a fusion of multiple techniques like patch resizing, attention related changes and CNN based modifications. These models had the highest F1-score among all the state-of-the-art models that have been reviewed in this paper. PiT model is similar to PoolFormer architecture, where the attention module is replaced by a pooling layer. But the accuracy of the PiT is better than PoolFormer because of the dimensionality related changes proposed in PiT (similar to CNNs). Here, the number of channels is increased as the layer progresses and the resolution of input image is reduced. Because of these changes to the dimensionality of the image, PiT has faster inference and training times compared to PoolFormer based model. Swin Transformer utilizes CNN based spatial reduction techniques. Similar to VOLO and TNT, the attention in Swin Transformer is calculated with respect to all the patches within a given window and shifted window as well as with the patches in the entire image. Hence Swin Transformer achieved the highest F1-score, when fine-tuned on APTOS dataset for DR detection, but the training and inference time was relatively higher compared to other transformer models. Twin Transformer is similar to Swin Transformer. The architectural changes to the Swin, made the Twin model lighter, easier and faster to train and infer. Twin model has the 2nd best F1-score when fine-tuned on APTOS dataset.

3 state-of-the-art Mixer architectures namely MLP-Mixer, ResMLP and Conv-Mixer, were evaluated and benchmarked. MLP-Mixer and ResMLP have similar architecture. ResMLP has an affine layer added to the MLP-mixer architecture, thereby improving the F1-score of the model. Vector handling in MLP-Mixer, decreased the training and inference time compared to ResMLP. Conv-mixer is a hybrid of ViT and MLP-Mixer, where images are divided into patches and passed through MLP-Mixer layers. F1-score achieved by Mixer models were lower than Transformer based architecture as the attention layer was replaced by MLP Layers and loss of global context. The training and inference times were also higher compared to other transformer based architectures. The performance of Mixer architectures was comparable to CNN based architectures in terms of F1-score, model size and training and inference times.

In summary 18 different state-of-the-art pre-trained models were fine-tuned and evaluated on APTOS dataset. These models included CNN, Transformers and Mixer based architectures. Global context results in higher F1-score, as it associates position of biomarkers with other anatomies of retina. Usually the biomarkers would be present in only 1% of the entire image, therefore capturing finer details is essential to improve the performance of the model. Certain transformer based models used global context as well as captured finer details, hence they yielded the best performance in terms of faster training speed, inference speed, better F1-score and smaller model size. The highest F1-score on APTOS dataset, as recorded in public leader-board and documented in literature was 85.6%. Most of the transformer based models consistently out-performed the state-of-the-art results. Among the Transformer based architectures Swin, Twin and PiT based models resulted in best performance. Hence, these three models were used for further evaluation. Swin model had the highest F1-score of 88.1%, but training time was almost 2.5 times more than the training time for Twin and PiT, Hence Swin model was not used for extensive ablation study. PiT was chosen for extensive ablation study as, it had similarities with both Swin and Twin models. Figure 4, indicates the F1-score for CNN, Transformer and Mixer architecture represented by ResNet, Swin and Res-MLP model respectively, with respect to the epochs.

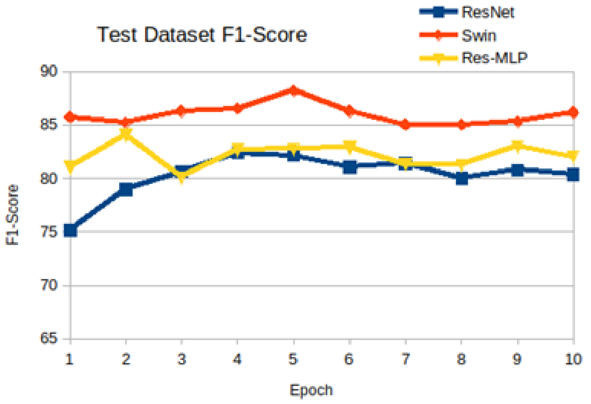

Figure 4. F1-scores CNN, Transformer and Mixer models

4.2 Ablation study

The PiT baseline model considered for extensive ablation study contained. Different experiments were conducted on the baseline model to identify the optimum hyperparameters. Table 3 shows the performance of the PiT model when the number of fully connected layers, normalization factors, dropout functions and batch sizes were experimented as part of ablation study. 2 fully connected layers seems to be optimum for APTOS dataset, as increasing or decreasing the number of fully connected layers in the classification-head decreases the F1-score. Changing the ImageNet normalization factor to default normalization factor to 0.5 also decreases the F1-score. Alpha dropout and Feature Dropout reduces the F1-score, when compared with the original dropout function with a probability of 30%. Batch size of 32, was the maximum value that could be considered because of the GPU capacity and the input image resolution. This was the batch size that was considered for the baseline PiT model. The other two batch sizes that were considered for evaluation were 16 and 8. It can be observed from the results that, a batch size of 32 leads to highest F1-score. The F1-score of the algorithm reduces as the batch size reduces, because when more examples are considered as a batch, the model generalizes better, thereby leading to a smoother gradient curve and faster inference. Time required for training is less, for higher batch sizes. 2 FC layers with ImageNet based normalization factor, normal dropout with probability of 30% and a batch size of 32 results resulted in the highest F1-score for APTOS dataset.

Different variants of ReLU based activation functions were experimented on the base PiT model and the results documented in Table 4. ReLU based activation function adds non-linearity as well as behaves like a natural dropout layer, hence it achieved the highest F1-score when compared to all other variants of ReLU for DR detection on APTOS dataset. RReLU is a variant of ReLU, where for the negative values of the input, the activations have a small negative slope rather than making it to zero. The value of the slope is chosen randomly, hence the performance of the algorithm also depends on the value of slope that has been initialized. Continuously Differentiable Exponential Linear Unit (CELU) and Scaled Exponential Linear Unit (SELU) have similar activation functions, and therefore F1-score is also similar. SiLU / Swish is a hybrid of sigmoid and ReLU based activations. Sigmoid based activation functions usually result in lower classification accuracy, hence the F1-score of SiLU is the least compared to all other activations. Gaussian Error Linear Unit (GELU) and Mish activation are similar to SiLU, but instead of sigmoidal function GELU utilizes Gaussian distribution and Mish uses the Tanh computations, thereby leading to higher F1-score.

Table 3. Initial ablation study on PiT model

32 batch, ImageNet norm, 2 FC, 30% dropout, ReLU, AdamW, no scheduler		
PiT Experiments	**Test F1-Score (%)**	**Train Time (min)**
Original PiT model	86.95	55.2
1 FC layer	86.11	54.5
3 FC layers	85.71	57.8
Normalize: 0.5	86.56	54.9
Alpha dropout	85.85	54.8
Feature Alpha Dropout	86.4	54.9
Batch 8	85.61	60.3
Batch 16	86.05	58.4

Table 4. PiT Transformer with different activations

PiT Experiments: 32 Batch, ImageNet Norm, 2 FC, Dropout, AdamW	Test F1-Score (%)
Rectified Linear Units, ReLU, (x)=max(0,x)	86.95
$RReLU(x) = \begin{cases} x\ , & if\ x > 0 \\ ax, & otherwise \end{cases}$	86.68
$SELU(x) = \lambda \begin{cases} x\ if\ x > 0 \\ \alpha e^{x} -\ \alpha, if\ x < 0 \end{cases}$	85.85
$CELU(x) = max(0, x) + min(0, \alpha * \left(exp\left(\frac{x}{a}\right) - 1\right))$	85.85
$SILU(x) = x * \frac{1}{1 + e^{-x}}$	85.3
$GELU(x) = x *\ \phi(x)$	85.85
$Mish(x) = x * Tanh(Softplus(x))$	86.81

Different optimizers were evaluated with the baseline PiT model with ReLU activations as shown in Table 5, to establish the most optimum optimizer for APTOS dataset. Learning rate for each of these optimizers was kept constant. Adam optimizer is a fusion of RMSprop and Momentum related concepts. Learning Rate (LR) in Adam is adaptive because of the component associated with RMSprop. Because of Momentum related components in Adam optimizer, it can accelerate and decelerate the learning process. Hence the performance of Adam optimizer is better than RMSprop. Nadam, AdamW, Adamax and Radam are updated versions of Adam optimizer, to enable better performance. Nadam is Adam optimizer with Nesterov momentum, thereby leading to better F1-score compared to Adam. In AdamW, the weight decay is decoupled from the gradient based update, thereby leading to better regularization and model generalization and performance compared to Adam optimizer. Adam's bad convergence is because, the adaptive learning rate has an undesirably large variance in the early stage of model training due to the limited amount of training samples used. RAdam is a variant of the Adam optimizer that seeks to tackle Adam's bad convergence problem by introducing a term to rectify the variance of the adaptive learning rate. RAdam optimizer leads to best F1-score compared with all other optimizers.

Table 5. PiT Transformer with different optimizers

PiT, 32 batch, ImageNet norm,2 FC, dropout, no scheduler,ReLU				
RMS-Prop	**Adam**	**NAdam**	**AdamW**	**RAdam**
86.12	86.26	86.67	86.95	87.64

Table 6. PiT experiments with different schedulers

32 batch, ImageNet norm, 2 FC, dropout, AdamW, ReLU			
PiT Experiments	**Test F1-Score (%)**	**PiT Experiments**	**Test F1-Score (%)**
Constant	86.54	Linear	85.86
Cos annealing	86.68	Cos annealing: warm restart	82.83
Multiplicative	86.81	Lambda	86.2
Exponential	86.81	No scheduler	86.95
Step	87.225	Multistep	87.36

Various learning rate schedulers were evaluated on the base PiT, to identify the best decay technique for DR detection, as shown in Table 6. The initial learning rate considered was 3*e-5. LR was increased using techniques like Constant LR scheduler and Linear LR scheduler. But the F1-score was lower because for image classification tasks, LR has to be decreased rather than increased. With cosine annealing scheduler techniques, LR decreases for first 10 epochs and then increases, in a cosine manner. Because the current model converged within the first 10 epochs, hence the performance was comparable to other techniques. But if the model would not have converged within the first 10 epochs, then the F1-score would have degraded. A modification of Cosine annealing scheduler was cosine annealing with warm restarts, where the LR increased initially and then dropped to zero (rather than in cosine fashion) and started to increase again. As the LR increased with this technique, the performance degraded. Lambda and multiplicative LR scheduler drastically reduces LR till it reaches saturation towards zero. But as the model converged by 5th epoch, hence the effect of learning rate drop was not reflected in the F1-score, otherwise it would have degraded the F1-score. Using step and multi step learning rate decay, the LR was reduced in steps and they have the best performance for DR detection on APTOS dataset.

In summary, for DR detection using APTOS dataset and PiT model, the best configuration was with a batch size of 32, ImageNet based normalization, 2 fully connected layers with dropout in the classification-head, ReLU activation, RAdam optimizer and “Multistep” based learning rate scheduler.

4.3 Performance on multiple opensource datasets

3 Transformer based models namely PiT, Swin and Twin were further evaluated and benchmarked on other opensource retinal fundus datasets as shown in Table 7. All the three

models had a batch size of 32, ImageNet based normalization, 2 fully connected layers with dropout, ReLU activation, RAdam optimizer, dropout and MultiStep LR decay function. It can be observed from Table 7, that all the 3 models that were considered for evaluation, consistently out-performed the state-of-the-art. For APTOS disease grading dataset, the state-of-the-art model (SOTA) [18] had the best F1-score of 86%, but all the models that were trained and evaluated in this paper, had an F1-score of above 86%. Swin model had the best F1-score of 88.7%. Using the selected methods on the base PiT model, seems to have increase the F1-score of PiT model to 87.95%.

Table 7. Experiments on various opensource fundus datasets

32 batch, ImageNet norm, 2 FC, dropout, AdamW, ReLU, multi-step				
Datasets	**Sl**	**Model**	**Test F1-Score (%)**	**Train / Epoch-Min**
APTOS Disease Grading Dataset	1	PiT	87.95	8.7
	2	Swin	88.7	23.6
	3	Twin	87.79	5.5
	4	SOTA: [18]	86%	-
APTOS Disease Detection Dataset	5	PiT	99.16	8.7
	6	Swin	99.30	22.3
	7	Twin	99.30	5.9
IDRiD-AMD	8	PiT	86.52	35.5
	9	Swin	85.52	61.4
	10	Twin	84.57	28.6
	11	SOTA: [10]	84	-
IDRiD-DR	12	PiT	64.16	35.6
	13	Swin	63.17	73.0
	14	Twin	63.25	31.8
	15	SOTA: [10]	63	-
AREDS	16	PiT	88.86	51.6
	17	Swin	90.53	103.3
	18	Twin	89.97	31.9
	19	SOTA: [12]	86	-
Messidor	20	PiT	84.87	25.6
	21	Swin	83.12	62.0
	22	Twin	85.25	19.1
	23	SOTA: [11]	73	-

Certain applications require the model to only detect if a patient has DR or not, in such cases it becomes a binary classifier. Hence the dataset is divided into two classes: normal class and disease class. For binary classification, the dataset contains almost similar number images. Swin, PiT and Twin Transformer models were fine-tuned and evaluated for binary classification. The results indicate very good F1-scores for all the three models, with more than 99% on the test dataset.

Apart from the APTOS dataset, the models were benchmarked on other opensource retinal fundus datasets like IDRiD-AMD, IDRiD-DR, AREDS and Messidor for the detection of DR and AMD. The state-of-the-art models for each of these datasets were reviewed and documented in Table 1. For IDRiD-AMD dataset, the state-of-the-art model [10] had the best F1-score of 84%, but all the models that were trained and evaluated in this paper, had an F1-score of above 84%. PiT model had the best F1-score of 86.5%. The state-of-the-art model [10] for IDRiD-DR dataset had the best F1-score of 63% when trained on IDRiD-DR and Eyepacs dataset and evaluated using IDRiD-DR dataset. But all the 3 models which were trained and evaluated in this paper, had an F1-score of above 63% when trained only on IDRiD-DR dataset without EyePacs dataset. PiT model had the best F1-score of 64.16%. For AREDS dataset, the state-of-the-art model [12] had an F1-score of 86%, where an ensemble of 6 models was created. But in this paper, each 3 of the individual models evaluated had F1-score of above 89% without the need to use an ensemble of models. Swin Transformer had the best F1-score for AREDS dataset of 90.53%. The state of the art model [11] for Messidor dataset achieved an F1-score of 73%. All the 3 models which were fine-tuned in this paper, had F1-score of above 83%, with Twin Transformer having the highest F1-score of 85.25%.

In summary, the models considered in this paper out-performed the state-of-the-art papers for all the datasets.

5. CONCLUSIONS AND FUTURE SCOPE

18 state-of-the-art pre-trained models based on CNN, Transformers and Mixer architectures were evaluated in this paper. The pre-trained models were coupled with a classification-head which contained 2 fully connected layers with dropout, activation layer and softmax layer. 6 datasets, 18 models, 7 activations, 5 optimizers and 10 LR decay methods were experimented. Overall a comprehensive analysis of over 71 experiments was conducted and a detailed analysis has been presented in this paper to select the most optimum models and methods for the detection of DR. The performance of Transformer based models outperformed CNN and Mixer based models for the detection of DR, because of the ability of the transformer to learn the global context and associate position of biomarkers with other anatomies of retina. Swin, Twin and PiT Transformers outperformed all other Transformer based models as well as state-of-the-art, because of their ability to encode fine as well as coarse-level details of biomarkers. Batch size of 32, ImageNet based normalization, ReLU activation, RAdam optimizer and multi-step LR decay were the most optimum methods for retinal disease detection.

Few of the potential applications of proposed DR detection framework includes early disease detection, eye screening using mobile health apps, screening for health insurance companies, educational tools for health care professionals and telemedicine.

Some of the limitations of the current study are: 1) To ensure uniformity in benchmarking the different models, augmentation techniques were not used to improve the performance of the model, 2) Cascade of models for detection of different retinal anatomies were not evaluated, because of the limited GPU capacity, 3) Models were not optimized and benchmarked for porting into edge device, 4) Real-time fundus capture and validation was not performed.

For improving the performance and usability of the proposed techniques, this research work can be extended to: 1) Implement augmentation techniques to increase the volume of the dataset, 2) Use ensemble of high performing classification models, 3) Fuse DR grading classifier, blood vessel and biomarker segmentation techniques into a single framework, 4) Validate on real-world fundus images, 5) Port the developed model onto edge device (Eg. mobile phones).

REFERENCES

[1] Krizhevsky, A., Sutskever, I., Hilton, G.E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929

[3] Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A. (2021). Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems, 34: 24261-24272.

[4] Pasha, L.T.M., Rajashekar, J.S. (2023). Diabetic Retinopathy severity categorization in retinal images using Convolution Neural Network. Revue d'Intelligence Artificielle, 37(4): 1031-1037. https://doi.org/10.18280/ria.370425

[5] Saleh, R.W., Abdullah, H.N. (2023). Early diabetic retinopathy detection using convolution neural network. Revue d'Intelligence Artificielle, 37(1): 101-107. https://doi.org/10.18280/ria.370113

[6] Dekhil, O., Naglah, A., Shaban, M., Ghazal, M., Taher, F., Elbaz, A. (2019). Deep learning based method for computer aided diagnosis of diabetic retinopathy. In 2019 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1-4. https://doi.org/10.1109/IST48021.2019.9010333

[7] Liu, S., Gong, L., Ma, K., Zheng, Y. (2020). GREEN: A graph residual re-ranking network for grading diabetic retinopathy. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, pp. 585-594. https://doi.org/10.1007/978-3-030-59722-1_56

[8] Zhuang, H., Ettehadi, N. (2020). Classification of diabetic retinopathy via fundus photography: Utilization of deep learning approaches to speed up disease detection. arXiv preprint arXiv:2007.09478. https://arxiv.org/abs/2007.09478

[9] Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M. (2021). Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors, 21(11): 3704. https://doi.org/10.3390/s21113704

[10] Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., et al. (2020). Idrid: Diabetic retinopathy–Segmentation and grading challenge. Medical Image Analysis, 59: 101561. https://doi.org/10.1016/j.media.2019.101561

[11] Seoud, L., Chelbi, J., Cheriet, F. (2015). Automatic grading of diabetic retinopathy on a public database. In Proceedings of the Ophthalmic Medical Image Analysis International Workshop, 2: 97-104. https://doi.org/10.17077/omia.1032

[12] González-Gonzalo, C., Sánchez-Gutiérrez, V., Hernández-Martínez, P., Contreras, I., Lechanteur, Y.T., Domanian, A., et al. (2020). Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. Acta Ophthalmologica, 98(4): 368-377. https://doi.org/10.1111/aos.14306

[13] Singh, K., Drzewicki, D. (2019). Neural style transfer for medical image augmentation.

[14] Wang, X., Lu, Y., Wang, Y., Chen, W.B. (2018). Diabetic retinopathy stage classification using convolutional neural networks. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 465-471. https://doi.org/10.1109/IRI.2018.00074

[15] Peng, Y., Dharssi, S., Chen, Q., Keenan, T.D., Agrón, E., Wong, W.T., Chew, E.Y., Lu, Z. (2019). DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology, 126(4): 565-575. https://doi.org/10.1016/j.ophtha.2018.11.015

[16] Dondeti, V., Bodapati, J.D., Shareef, S.N., Naralasetti, V. (2020). Deep convolution features in non-linear embedding space for fundus image classification. Revue d'Intelligence Artificielle, 34(3): 307-313. https://doi.org/10.18280/ria.340308

[17] Nasir, N., Afreen, N., Patel, R., Kaur, S., Sameer, M. (2021). A transfer learning approach for diabetic retinopathy and diabetic macular edema severity grading. Revue d'Intelligence Artificielle, 35(6): 497-502. https://doi.org/10.18280/ria.350608

[18] Chaturvedi, S.S., Gupta, K., Ninawe, V., Prasad, P.S. (2020). Automated diabetic retinopathy grading using deep convolutional neural network. arXiv preprint arXiv:2004.06334. https://arxiv.org/abs/2004.06334

[19] Sheikh, S.O. (2020). Diabetic reinopathy classification using deep learning. Master's thesis.

[20] Kassani, S.H., Kassani, P.H., Khazaeinezhad, R., Wesolowski, M.J., Schneider, K.A., Deters, R. (2019). Diabetic retinopathy classification using a modified xception architecture. In 2019 IEEE international symposium on signal processing and information technology (ISSPIT), pp. 1-6. https://doi.org/10.1109/ISSPIT47144.2019.9001846

[21] Wang, L., Schaefer, A. (2020). Diagnosing diabetic retinopathy from images of the eye fundus. Cs230. Stanford. Edu.

[22] Patel, R., Chaware, A. (2020). Transfer learning with fine-tuned MobileNetV2 for diabetic retinopathy. In 2020 international conference for emerging technology (INCET), pp. 1-4. https://doi.org/10.1109/INCET49848.2020.9154014

[23] Pour, A.M., Seyedarabi, H., Jahromi, S.H.A., Javadzadeh, A. (2020). Automatic detection and monitoring of diabetic retinopathy using efficient convolutional neural networks and contrast limited adaptive histogram equalization. IEEE Access, 8: 136668-136673. https://doi.org/10.1109/ACCESS.2020.3005044

[24] Xie, L., Vaghefi, E., Yang, S., Han, D., Marshall, J., Squirrell, D. (2023). Automation of macular degeneration classification in the AREDS dataset, using a novel neural network design. Clinical Ophthalmology, 455-469. https://doi.org/10.2147/OPTH.S396537

[25] Bodapati, J.D., Shaik, N.S., Naralasetti, V. (2021). Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. Journal of Ambient Intelligence and Humanized Computing, 12(10): 9825-9839. https://doi.org/10.1007/s12652-020-02727-z

[26] Mutawa, A.M., Sruthi, S. (2022). Diabetic retinopathy classification using vision transformer. In 2022 6th European Conference on Electrical Engineering & Computer Science (ELECS), pp. 25-30. https://doi.org/10.1109/ELECS55825.2022.00012

[27] Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J. (2021). Vision Transformer-based recognition of diabetic retinopathy grade. Medical Physics, 48(12): 7850-7863. https://doi.org/10.1002/mp.15312

[28] Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K. (2021). Is it time to replace cnns with transformers for medical images?. arXiv preprint arXiv:2108.09038. https://arxiv.org/abs/2108.09038

[29] Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., Wang, Z. (2022). FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. Frontiers in Physiology, 13: 961386. https://doi.org/10.3389/fphys.2022.961386

[30] Dihin, R.A., Al-Jawher, W.A.M., AlShemmary, E.N. (2022). Diabetic retinopathy image classification using shift window transformer. International Journal of Innovative Computing, 13(1-2): 23-29. https://doi.org/10.11113/ijic.v13n1-2.415

[31] Jin, X., Li, H., Li, R. (2022). Dformer: Dual-Path transformers for geometric and appearance features reasoning in diabetic retinopathy grading. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 401-416. https://doi.org/10.1007/978-3-031-18910-4_33

[32] He, K., Zhang, X., Ren, S., Sun, J. (2016). Identity mappings in deep residual networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

[33] Tan, M., Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pp. 6105-6114.

[34] Bao, H., Dong, L., Piao, S., Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254. https://arxiv.org/abs/2106.08254

[35] Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S. (2022). Volo: Vision outlooker for visual recognition. IEEE Transactions on Pattern Analysis and Machine intelligence, 45(5): 6575-6586. https://doi.org/10.1109/TPAMI.2022.3206108

[36] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y. (2021). Transformer in transformer. Advances in Neural Information Processing Systems, 34: 15908-15919.

[37] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pp. 10347-10357.

[38] Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., Tian, Q. (2021). Visformer: The vision-friendly transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 589-598.

[39] Dai, Z., Liu, H., Le, Q.V., Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems, 34: 3965-3977.

[40] El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., Jégou, H. (2021). XCIT: Cross-covariance image transformers. Advances in Neural Information Processing Systems, 34: 20014-20027.

[41] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H. (2021). Going deeper with image transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 32-42.

[42] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J.S., Yan, S.C. (2022). Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10819-10829.

[43] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., Jégou, H. (2022). Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 5314-5321. https://doi.org/10.1109/TPAMI.2022.3206148

[44] Trockman, A., Kolter, J. Z. (2022). Patches are all you need?. arXiv preprint arXiv:2201.09792. https://arxiv.org/abs/2201.09792