Using Data-Efficient Image Transformers for Diabetic Retinopathy Severity Classification

Veda Fernandes [*†]

October 17, 2023

Abstract

Roughly 10% of the global adult population is diabetic, diabetes is a metabolic condition which results in chronically high blood sugar levels. Patients with diabetes are at substantially higher risk for several serious health conditions including diabetic retinopathy (DR). DR is a vision-threatening disease which affects 35% of diabetic patients and is projected to affect 160 million people by 2045. Diabetic patients should be screened for retinopathy every one to two years; however, in many countries patients are not regularly screened and therefore not treated. Globally, the lack of rapid and cost-effective screening strategies for DR leads to underdiagnosis and loss of vision. Machine learning tools offer a solution in developing automated models to diagnose DR from eye fundus images. In published literature, convolutional neural networks (CNNs) are the state-of-the-art model for classification of DR. More recently, transformer models have been applied and shown superior performance. Text transformer models have resulted in the proliferation of tools such as ChatGPT, which provide contextual understanding and ability to identify dependencies. In this study, we perform a head-to-head comparison between CNN and vision transform models for classifying DR. We demonstrate that transformer models diagnose DR with a substantially higher accuracy, ranging up to 13% as measured by the F1 performance metric. Furthermore, we identify optimal training parameters for diagnosis of DR, training a total of 19 machine learning models reaching a test set F1 score performance of 90% on a dataset of 35,130 fundus images with 20% of images withheld for independent testing.

1 Introduction

Diabetic retinopathy (DR), is a vision-threatening microvascular disease caused by significant damage to the blood vessels in the retina and is one of the most frequent complications of diabetes mellitus [NPS22]. DR is a leading cause of

[*] Dubai International Academy Emirates Hills

[†] Advised by: Dr. Parsa Akbari, University of Cambridge

preventable blindness and vision impairment among the working-age population, with a prevalence of about 35% among those with diabetes mellitus. By 2045, it is estimated that 783 million people will be diabetic [Fed21] and 160 million people could be affected by DR [TTY+21]. The prevalence increase of DR by 2030 is notably concentrated in the low and middle income countries in regions such as Asia, South America and the MENA. [TW22]. This disease burden will require effective DR screening strategies to align with the changing demographic. About 56% of new cases could be reduced with timely monitoring of severity and treatment [TMS20].

DR can be graded into 5 stages according to morphological changes that occur in the retina as the disease progresses: non-proliferative diabetic retinopathy (NPDR), mild NPDR, moderate NPDR, severe NPDR and Proliferative Diabetic Retinopathy [DSS17] (Fig. 1). Screening is imperative to identify the stage of DR - with timely referral, the progress of DR can be slowed and severe vision-impairment can be prevented [WSK+18]. Despite this, there is a shortage of ophthalmologists to screen millions of retinal images for each diabetic patient, especially in developing countries. [RLW+20] [WSK+18]. Here, technology can provide an alternate solution to traditional screening methods by physicians by reducing the cost and manpower required to screen patients for PDR.

Ocular telemedicine is a concept which has been proposed to make DR screening more cost efficient. This involves local clinics sending images of the retina to a central 'grading center' where experts can grade the level of DR severity [HSCA16]. Hand-held imaging devices are another solution and have achieved high specificity and sensitivity compared to traditional retinal cameras [PDKB22]. However, in both solutions, trained clinical professionals are still required to analyze the retinal images. An automated system would conduct the initial screening of retinal images to detect signs of DR even when ophthalmologists are unavailable.

An automated system must recognize the changes to retinal vasculature caused by the various stages of DR as seen on fundus images. As seen in Fig. 1, DR affected retinal images show characteristic color and patchy variations on the fundus image, due to morphological changes in the retina. These changes include lesions called MicroAneurysms (MA), Hard EXudates (HEX), Soft Exudates (SE), HEMorrhages (HEM), and an increase in blood vessels. MAs are seen in one quadrant in mild NPDR, and it progresses to vessel blockage, and presence of lesions in moderate NPDR. Severe NPDR presents with venous beading and a large number of HEMs. This is a precursor to 'neovascularization' of the PDR stage [NPM+22], or to the formation of new blood vessels on the retina, which may eventually lead to blindness (Fig. 1).

In recent decades, artificial intelligence has been trained to classify DR stages from fundus images. Early models used ML-based classifiers like Random Forest (RF) ([CSC+14], K-Nearest Neighbours (KNN) [NvGRA07] , Support Vector Machines (SVM) [SAFL10] and Artificial Neural Networks (ANN) [UDH+04]. These methods required efficient prior hand-engineered feature extraction, which could introduce errors into complex fundus imaging. [NG18] evaluated 7 automated retinal image analysis (ARIA) systems to classify DR. These models had

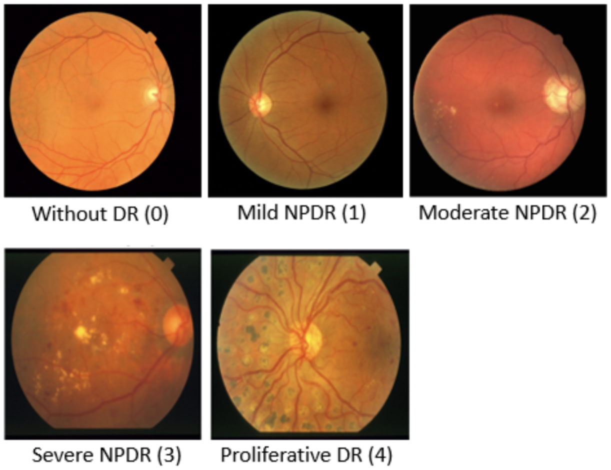

Figure 1: Images of the retina showing stages of diabetic retinopathy, graded into 5 classes ranging from 0-4, as per the EyePACS dataset [GPC+16].

sensitivities of 87%-95%, but had limited specificities of 50%-69%, thereby leading to a high number of false positives that negatively impacted the clinical application of these systems [NG18].

Deep learning (DL), also known as deep neural networks, is an emerging field which relies on multiple layers of processing to extract high level features from input images. The application of DL was previously limited by the availability of computational processing power. However recent innovations including cheaper compute costs, cloud computing, and specialized ASIC's which are processing units optimized for ML, have made DL more broadly available. DL has a wide range of applications from computer vision and finance to autonomous vehicles [DDD20]. It has also been applied in medicine and in recent years has produced favorable results in the diagnosis of DR in fundus imaging, with sensitivities of 80.28% to 100.0% and specificities of 84.0% to 99.0% [NLA+19].

Convolutional Neural Networks (CNNs) are popular in deep learning for image classification tasks, including medical imaging [BR18]. The design of CNNs maximizes their efficiency by focusing on a smaller section of the data; however, this compromises their performance when capturing broader patterns and relationships across spaced out parts of an image [SKZ+23].

A CNN of the Inception-v3 architecture [GPC+16] and CNN-based residual learning [GL17] were used to detect DR. To classify fundus images into DR severity stages, the literature shows several CNN architectures were used. A CNN was used to identify exudates [YXK17], transfer learning and hyperparameter tuning was done on the CNNs of AlexNet, VGGNet, GoogleNet and ResNet [WLZ18], a DenseNet architecture with hyperparameter tuning was trained [RPC+20] and an attention mechanism coupled with a modified DenseNet-169 architecture was used [FFAH22].

Recently, transformer models have gained popularity in a variety of applications. An example of this rising interest in transformers is ChatGPT - an ML-tool which uses a transformer architecture to do NLP [Ope23]. Transformers are efficient in identifying and understanding the relationship between separate elements within data and are capable of parallel processing, which means they are able to be trained more rapidly [IEE$^+$23]. In addition to text, the success of transformers in Natural Language Processing (NLP) prompted the creation of vision transformers (ViT) which can be applied to image data to carry out computer vision tasks [HGL$^+$23]. These traits make ViTs adept at doing tasks that require an understanding of the context, thus making them valuable not only in the field of NLP but also in medical imaging classification and segmentation [SKZ$^+$23]. Consequently, ViTs may be a prospective model architecture for DR detection through fundus images.

[WHX$^+$21] and [AAKJTT$^+$21] demonstrated that attention-based ViTs provide high accuracy for DR classification. [AKCS23] used an ensemble of Vision Transformers (ViT), Data efficient image Transformers (DeiT), Bidirectional Encoder representation for image Transformer (BEiT) and Class-Attention in Image Transformers (CAIT) to stage DR severity. Transformers gained prominence only recently and hence, there is limited literature on the effect of hyperparameters of ViTs as compared to CNNs, which have been studied much more rigorously. Therefore, this study aims to compare the performance of a ViT and CNN architecture and find the optimal hyperparameters for a ViT model to classify DR fundus images.

In this study, we benchmarked DeiT, a recent ViT model, against ResNet-18, a classical CNN model, for binary classification of retinal fundus images into no or mild nonproliferative DR and moderate nonproliferative or more severe DR. This classification was chosen as it is recommended by the American Diabetes Association and the International Council of Ophthalmology that cases of moderate NPDR or more severe stages are referred to an ophthalmologist [WSK$^+$18] [SCD$^+$17]. Therefore, the model would be used as a preliminary screening tool to help refer patients while the specific diagnosis of the severity of DR could be done by a medical professional upon consultation. We used the EyePACS Dataset, consisting of 35,130 fundus images [GPC$^+$16].

To address the disparity in literature for the application of the transformer models in DR classification, we identified the optimal hyperparameters for the DeiT architecture and compared its performance to ResNet-18. The models were evaluated on their F1 scores, which is a metric that measures the harmonic mean of the precision and recall of the model, penalizing any extreme values from either [HST$^+$22]. The results indicated that the transformer models showed superior performance across a range of hyperparameters including learning rate and batch size. The F1 scores showed that the DeiT model performed considerably better than the ResNet-18 model and that a learning rate of 1E-04, batch size 32, and epoch of 6 gave the best model performance.

2 Results

2.1 DeiT model performance showed 23% improvement in F1 score with optimal learning rate

We investigated the impact of learning rates on the performance of two model architectures - DeiT and ResNet-18. DeiT is a vision transformer while ResNet-18 is a convolutional neural network. The learning rate governs how quickly a model learns. A higher learning rate allows for the model to take larger steps to improve, however this may result in the optimal solution being overshot. In contrast, smaller learning rates may eventually reach a desirable result but are less time-efficient. To find the optimal value for the learning rate, each model was trained with 5 learning rates: 1E-03, 1E-04, 1E-05 and 1E-06. Both the ResNet-18 model and the DeiT model showed a similar trend in performance based on learning rates (Fig. 2).

For the DeiT model, learning rates of 1E-04 and 1E-05 demonstrated superior performance showing 40% improvement in test F1 score compared to higher learning rates. However, learning rates lower 1E-04 and 1E-05 deteriorated the ability of the model to detect moderate NPDR or more severe cases with a 7% decrease in model performance when tested with a learning rate of 1E-06. For the ResNet-18 model, similar to the DeiT model, the learning rates of 1E-04 led to better overall model performance. Below and above those values, the model performance was much poorer.

A significant finding was that the DeiT model consistently outperformed ResNet-18 across all learning rates (Fig. 2), with a 13% higher Test F1 score. ViT generally excels in understanding the bigger context in images as compared to CNN models, which could explain the results.

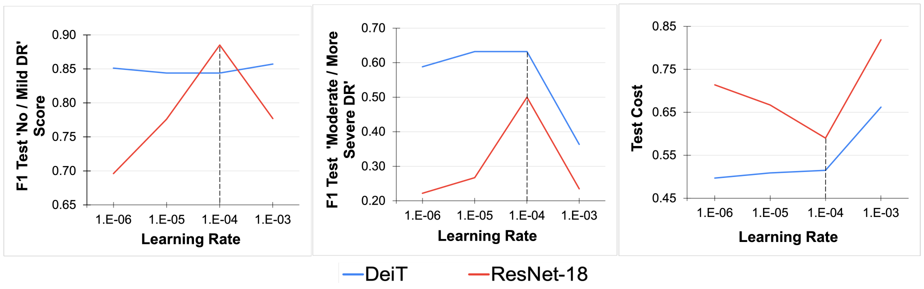

Figure 2: Graphs showing the effect of different learning rates on the performance for the DeiT and ReNet-18 models. This figure shows the Test F1 scores for the no or mild NPDR and moderate NPDR or more severe DR categories and overall cost of the DeiT and ResNet-18 models. The DeiT model generally performs better than the ResNet-18 model, with the F1 score being higher on average and the precision being lower. It can also be seen that the optimal learning rate for both models is 1E-04 as the F1 graphs peak while the cost function is low at that point.

2.2 Batch size of 32 improved model test F1 score by 50%

Deep learning models are trained with the stochastic gradient descent algorithm which performs each iteration of training using a single batch of data. Therefore, at each training iteration, improvements in model performance are incremental and dependent only on images present in the batch. Batch size is a critical training parameter which determines the number of images the model is trained or tested on in each iteration. Finding the correct batch size is important, as this fundamentally affects the training of the model. Larger batches result in a greater amount of information informing each training iteration; however, larger batch sizes are computationally intensive and are slow to execute. Smaller batch sizes are computationally efficient; however, less information informs each training iteration. Furthermore, smaller batch sizes lead to fluctuations in model training which may be beneficial in searching the model parameter space and resulting in superior performance, or may lead to ineffective training iterations and poor performance.

We experimented with a range of values, starting at 8 and increasing in factors of 2, testing batch sizes of 8, 16, 32, and 64 to assess their impact on the model's performance. We found that larger batch sizes tended to improve model performance, with a batch size of 32 giving the best overall performance (Fig. 3). While increasing batch size to 64 increased the model's ability to identify no or mild DR, it negatively impacted its ability to diagnose moderate NPDR or more severe DR by 10%, which is counterproductive to the aim of this model.

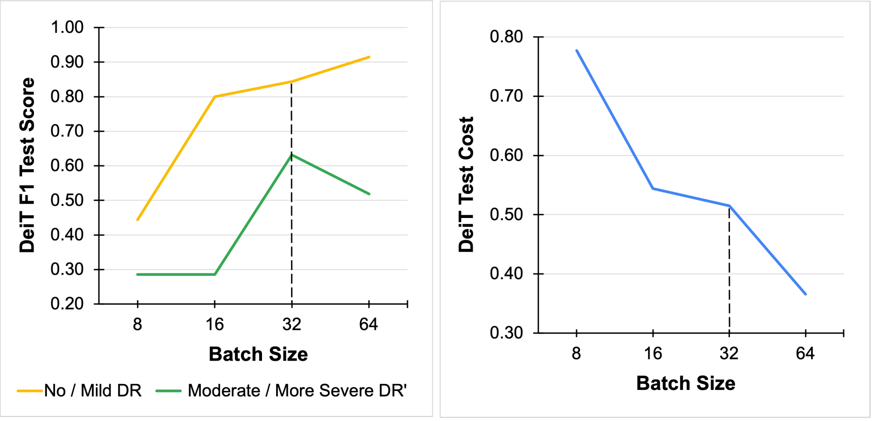

Figure 3: Graph showing the effect of different batch sizes on the performance of the DeiT model. This figure shows the Test F1 scores for the no or mild NPDR and moderate NPDR or more severe DR categories and overall cost of the DeiT and ResNet-18 models. The DeiT model generally performs better than the ResNet-18 model, with the F1 score being higher on average and the precision being lower. It can also be seen that the optimal learning rate for both models is 1E-04 as the F1 graphs peak while the cost function is low at that point.

2.3 Epoch of 6 improved test F1 score by 7%

Epochs control the number of times the model iterates through all the training fundus images, where a single iteration over all images in the training set is one epoch. Training the model across multiple epochs often results in superior performance as the model parameters are further improved from the training examples. The intention of model training is for the model parameters to be tuned to detect patterns in the training set which will be predictive of DR in future fundus images. However, training across too many epochs will result in overfitting, because the resulting model is over-adjusted for the training set and identifies patterns which are specific to the peculiarities of the training set but do not generalize to new data. Overfitting is detected by assessing the model on a testing set which is not utilized for model training.

An ideal number of epochs avoids overfitting or underfitting by having either too large or too small a number of epochs. To determine the optimal number of epochs, the DeiT model was tested on epochs of 2, 4, 6, 8 and 10. There was no significant relationship between the number of epochs and model performance, but the model achieved the highest average test F1 Score of 90% and 70% for each test class with 6 iterations (Fig. 4). The model performance decreased by an average of 7%, above and below 6 epochs.

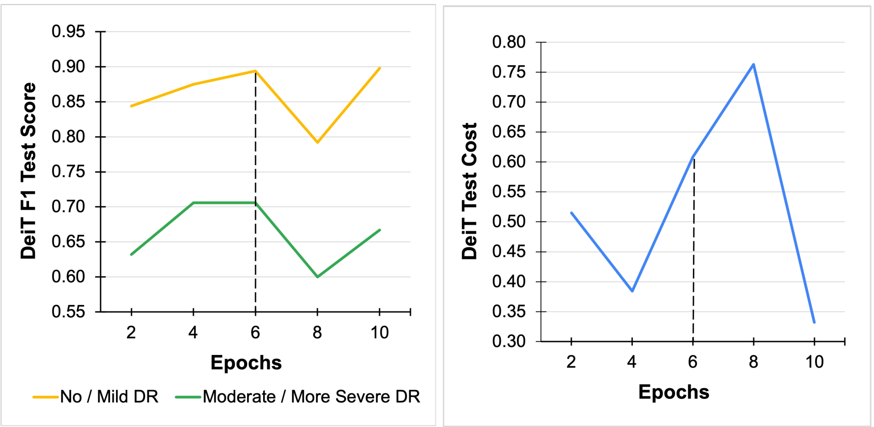

Figure 4: Graph showing the effect of different epochs on the performance for the DeiT model. This figure shows the Test F1 scores for no or mild NPDR and moderate NPDR or more severe DR categories and overall cost of the DeiT models. The optimal epoch was identified as 6 epochs as it showed the best F1 and cost results.

3 Discussion

In the study we have demonstrated that vision transformer models show superior predictive performance in diagnosis of DR compared to classical CNNs, with a 13% higher test F1 accuracy. We identified the optimal values for the DeiT model parameters such as learning rate, batch size and number of epochs. We analyzed the test F1 score for 19 machine learning models and found that hyperparameters learning rate of 1E-04. Additionally, a larger number of epochs of 6 and larger batch sizes of 32 proved to show better model performance. The DeiT model effectively classified images into both classes. A challenge that we encountered was the imbalance in the number of images in each category in the EyePACS dataset. There were over 22,000 images in the no or mild NPDR category and approximately 5400 images within the moderate DR or more severe DR category. Although the classes were weighted to reduce the effect of the imbalance, there was still a significant difference in the ability of the models to classify the images - the model performed 20% better at classifying images into no DR or mild NPDR than the moderate or more severe cases. However, overall, the DeiT model performed better than the ResNet-18 model, demonstrating the promise of vision transformers in DR classification applications. DeiT models can potentially reduce the need for manual screening.

Previous works largely focused on CNNs, with only a few papers on the applications of ViTs to DR classification [WHX^{+}21] [AAKJTT^{+}21] [AKCS23] . Recent research has shown that transformers provide significant advantages compared to CNNs, including providing contextual understanding and making connections between disparate features in the input image [MDB23]. However, there have not been direct comparisons between the performance of ViT and CNN models for diagnosis of DR. In this study, we aim to bridge this gap by doing a comprehensive comparison between a classical CNN and a recent ViT model.

Expanding on the work in this paper, DeiT models can be applied to multi-class DR classification tasks based on the severity of clinical symptoms. However, to accomplish this work with high accuracy and to improve the feature extraction capabilities, a larger dataset can be used, with a more balanced number of images for each class. More preprocessing techniques can be explored to arrive at better convergence of the loss function. The high imbalance in classes can be further addressed by custom data augmentation techniques such as rotation, adjusting brightness, contrast etc. of the images [AKCS23]. Additionally, variable illuminations and saturations of the images are a barrier to accuracy of predictions. Luminosity normalization is a pre-processing technique that could be applied to the images to improve the problem of variable illuminations. To deal with zero pixels, using the generic cropping functions may cause a loss of image data and disturb fundus geometry. Using a custom cropping window of variable lengths depending on the resolution of the images will preserve crucial information which will help in convergence of the loss function [RPC^{+}20].

4 Methods

The Pytorch Package in the Visual Studio Code environment was used to run the models. The experiment was run on fundoscopy images from the EyePACS database using two deep learning networks - DeiT, which is a vision transformer and ResNet-18, which is a convolutional neural network.

Fundoscopic examination is a routine clinical examination of the retina using an ophthalmoscope, and is used to detect numerous eye diseases including diabetic retinopathy. EyePACS is a telemedicine healthcare provider which offers diabetic retinopathy screening solutions in the United States. The EyePACS dataset includes 5 million fundoscopy images of a diverse population of healthy patients and those with various stages of diabetic retinopathy [GPC$^+$16]. 35,130 EyePACS color fundus images were utilized for the analysis, with 28,102 images set for training and 7,028 for testing. We ensured that all pairs of eyes from a single patient were kept together in the training and testing sets to ensure the testing set is independent from the training set.

While the demographic variables for the EyePACS dataset are not published, an EyePACS dataset of 9963 images from 4997 patients used in a paper had an average age of 54.1 (± 11.3) years, with 62.2% women [GPC$^+$16] The EyePACS fundus images are labelled in five classes as Normal (0), mild (1), moderate (2), severe (3), and proliferative DR (4) as shown in Fig. 1. Each image was rated by a clinician for the stage of DR present according to the International Clinical Diabetic Retinopathy severity scale (0-4) [GPC$^+$16]. For our model, we divided these classes into 2 categories based on recommended screening strategies [WSK$^+$18] [SCD$^+$17].

The performance of a deep learning model will be dependent on the quantity and integrity of the dataset used for training the model [ZLLS21]. The EyePACS dataset images are of various sizes and have pre-existing image noise and inconsistencies including the presence of artefacts, images being out of focus, underexposed or overexposed [TPM23]. The images were acquired by various cameras supported by the EyePACS platform, with field views of $40^\circ - 50^\circ$ [KOM$^+$20] and have different resolutions [TPM23]. There was no standard orientation of the images and they could be inverted as well, making it difficult to tell left from right eye images.

To improve the accuracy and reduce error rates of deep learning networks, a set of operations for pre-processing the images was required to train the model [ZLLS21]. Initially, the data was pre-processed by first converting them to tensors, then resizing the images to a 224 x 244 resolution and finally using the Random Horizontal Flip transformation to add more variance to the dataset.

Training deep learning models for high performance requires efficient optimization of the hyperparameters of the model. First we trained both DeiT and ResNet-18 models across learning rates between 1E-06 to 1E-03. Increasing the number of epochs or the batch size will ensure the model is adequately trained, thereby converging to the optimal solution and improving the accuracy. The DeiT model performed consistently better than the ResNet-18 model, over a range of learning rates, thus the DeiT model was chosen for further experimentation. We varied the batch size of the DeiT model from 4 to 64 and epochs from 2 to 10. The run time for the model varied due to increasing the number of epochs and decreasing the learning rate, which made the model take longer to train. The metrics, including F1, precision and recall, were logged for both the training and testing data. We generated a summary of the overall performance of the model under various hyperparameters to be visualized as graphs. We ran 13 different trials, making changes to the stated hyperparameters to identify which gave the best model performance. Every three iterations, the model performance was evaluated using the testing dataset of 7027 images.

4.1 Residual Neural Networks (ResNets)

Experimental work has shown that deeper networks are crucial for better performance but are more difficult to train and accuracy gets saturated beyond a point. Residual blocks offer a solution to the problem. First, the skip connection sets a short-cut to backpropagate the gradient as shown in Fig 5 . Second, due to the skip connection, instead of fitting the original identity mapping $H(x)$, the stacked layers need to fit or learn only the residual mapping of $F(x) = H(x) - x$.

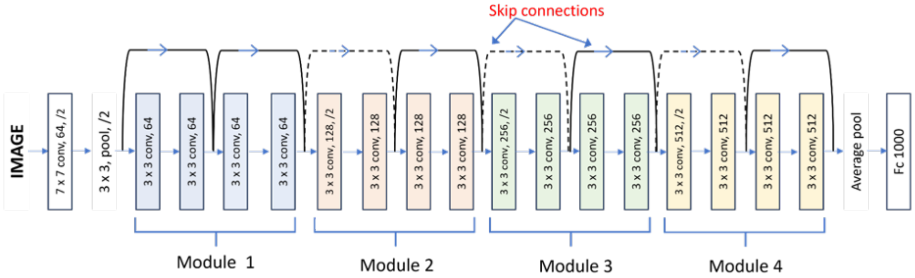

Figure 5: ResNet-18 architecture with skip connections in the identity blocks and convolutional blocks. Each layer consists of identical convolutional networks. When the input and output dimensions differ, convolutional skip connection is used, indicated in dotted lines.

If the identity mapping $H(x) = x$ is the desired underlying mapping, the residual mapping becomes $F(\mathrm{x}) = 0$, which makes the learning process easier - the weights and biases of the upper weight layer have to be pushed to zero. Residual blocks can forward propagate faster due to the skip connections [ZLLS21]. Thus, $H(x) = F(x) + x$, adding the output from previous layers to later layers. $y_i = F(x_i, W_i) + x_i$ represents each block, with x_i and y_i representing the input and output vectors of the layer. $x_{i+1} = f(y_i)$ where f denotes a ReLU activation function. F is a residual function [HZRS16]. The ResNet-18 architecture used in this paper has 18 layers. The first 2 layers are a 7×7 convolutional layer with 64 output channels and a stride of 2 , followed by a 3×3 max-pooling layer with a stride of 2 . ResNet-18 has 4—residual block modules. Each residual block has two layers, making $F = W_2 \sigma(W_1 x)$, where σ denotes a ReLU function (Fig. 6). Each block consists of 2 convolution layers, batch normalization, and corresponding ReLU activations. The convolutional layers have 3×3 filters. When the feature map size is halved, the number of filters is doubled, and down-sampling is with a stride of 2. The end of the network is a 1000-way fully connected layer with softmax and an average pooling layer. The shortcut connections are introduced to each pair of 3 × 3 filters as shown in Fig. 5 [HZRS15].

The standard ResNet residual block is called the Identity Block (Fig. 6). It is modified into the Convolutional Block when the input activation does not have the same dimension as the output – usually 1 X 1 convolutions are done, with a stride of 2 to match dimensions (Fig. 7) [HZRS15].

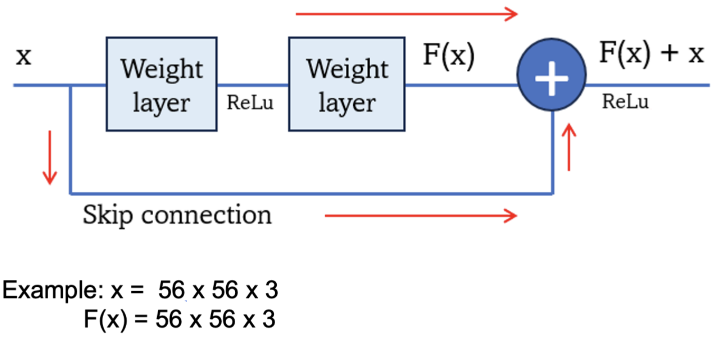

Figure 6: Identity Block - When input and output dimensions are equal, no additional layer in skip connection path

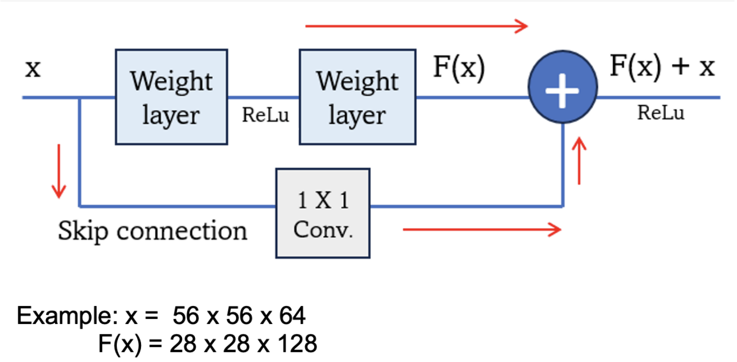

Figure 7: Convolutional Block - When input and output dimensions are not equal, 1 × 1 convolutional layer added in skip connection path

4.2 Data-efficient Image Transformers (DeiT)

Transformer models were initially applied in natural language (text) processing applications using an 'attention' mechanism which allowed modelling of elements within sentences of text without regard to their distance in the sequence [VSP+17]. Vision transformers applied these concepts to image processing and classification [DBK+20]. However, compared to classical CNNs, ViT models depend on pre-training using large amounts of data. Data-efficient image transformers aim to address the requirement for large amounts of training data by using a knowledge distillation technique to train a modified transformer, built on the ViT architecture proposed by [DBK+20], which transfers knowledge from a larger CNN to a smaller model. This reduces the training requirement to about 3 days and also requires less infrastructure. [TCD+20]

DeiTs are image transformers that propose a novel distillation procedure built upon the transformer block proposed by [DBK+20] and are trained on the ImageNet dataset only [TCD+20] . This contrasts with ViT which needs pre-training on hundreds of millions of images of curated data from many datasets showing high performance [DBK+20]. DeiTs are an effective method as they require lower volume of data and memory footprint for a given accuracy [AKCS23], making them a less computationally expensive architecture.

The knowledge distillation procedure using attention is the central principle of DeiT where the transfer of knowledge happens from 'the teacher' model to the 'the student' model [TCD+20]. The 'teacher' is RegNet Y–16GF, a CNN pre-trained on ImageNet. The student is a modified ViT architecture where the output of the 'teacher' is passed as an input to the 'student'.

In DeiT, a new distillation procedure is introduced where the teacher's hard decision is taken as the true label. DeiT decomposes each RGB image into a series of N patch tokens of 16 × 16 pixels each and converts it into a linear layer of 16 × 16 × 3 = 768 dimensional representation. A new distillation token is included which interacts with the class and patch tokens through the stack of transformer encoder layers (Fig. 8). The encoder layers contain Multi-head Self Attention (MSA) and Feed Forward Network (FFN) modules [TCD+20] [DBK+20]. The hard decision is illustrated below.

Let $y_t = \operatorname{argmax}_c Z_t(c)$ be the hard decision of the teacher. The task of the distillation token is to reproduce the hard decision y_t predicted by 'the teacher' and the class token has to reproduce the true label y. DeiT's loss function is given by:

$$\mathcal{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2}\mathcal{L}_{\text{CE}}\left(\psi\left(Z_s\right), y\right) + \frac{1}{2}\mathcal{L}_{\text{CE}}\left(\psi\left(Z_s\right), y_{\text{t}}\right)$$

Z_s and Z_t are the logit functions of the student and teacher models. ψ is the softmax function, and L_{CE} is the cross-entropy loss. Distillation tokens and the class token learn by back propagation and the distillation allows the model to learn from the teacher output [TCD+20], and this process is more efficient and requires less computational power than other vision transformers.

Following the training and testing of the chosen DeiT model, which shows its efficient performance in classifying diabetic retinopathy fundus images, it holds promise for application in telemedical or national screening programs to screen for severity of DR.

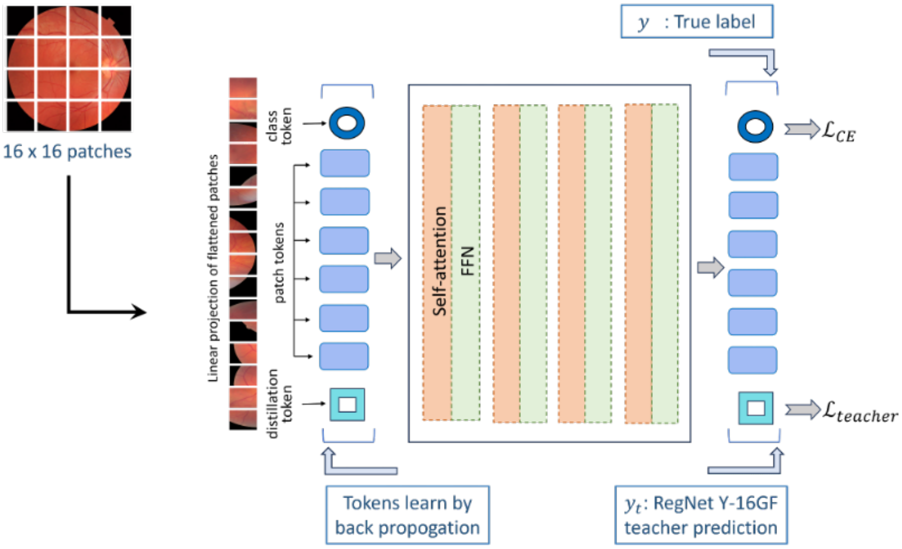

Figure 8: DeiT distillation procedure - The distillation token interacts with the class and patch token through the transformer encoders. The encoders in DeiT consist of repeated layers of self-attention and feed-forward network (FFN) blocks. The objective of the distillation token is to reproduce the teacher's prediction instead of the true label. The distillation and class tokens learn by back propagation.

4.3 Performance Metrics

The F1 score is a measure of a model's accuracy which takes into account both precision and recall. Precision and recall calculate the percentage of diagnoses which are correct and the percentage of DR patients which are diagnosed respectively. Furthermore, we calculate the F1 metric per outcome category and report all independently. The purpose of utilizing the F1 metric is to consider both true positive and false negative results by the model.

$$F_1 = \frac{2 \times \text{ precision } \times \text{ recall}}{\text{precision } + \text{ recall}}$$

5 Conclusion

The incidence of diabetes in the global population has reached 10%, patients with diabetes are at high risk for diabetic retinopathy and should be tested for DR every one or two years according to standard clinical guidelines. The lack of rapid and cost effective methods for diagnosis of DR is a major limiting factor for providing appropriate patient care [WWC+22]. Our study has shown that recent vision transformer methods have superior performance to a CNN model for the classification of DR. We have identified optimal model training parameters for the DeiT architecture. Our work demonstrates the ability of ViTs to improve the accuracy of automated DR classification.

References

[AAKJTT$^+$21] Nouar AlDahoul, Hezerul Abdul Karim, Myles Joshua Toledo Tan, Mhd Adel Momo, and Jamie Ledesma Fermin. Encoding retina image to words using ensemble of visiontransfo rmers for diabetic retinopathy grading. *F1000Research*, 10, September 2021.

[AKCS23] Chandranath Adak, Tejas Karkera, Soumi Chattopadhyay, and Muhammad Saqib. Detecting severity of diabetic retinopathy from fundus images using ensembled transformers. *arXiv*, 2023.

[BR18] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47, August 2018.

[CSC$^+$14] Ramon Casanova, Santiago Saldana, Emily Y Chew, Ronald P Danis, Craig M Greven, and Walter T Ambrosius. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS One*, 9(6), June 2014.

[DBK$^+$20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, October 2020.

[DDD20] Chitra A. Dhawale, Kritika Dhawale, and Rajesh Dubey. A review on deep learning applications. In *Advances in Systems Analysis, Software Engineering, and High Performance Computing*, pages 21–31. IGI Global, 2020.

[DSS17] Elia J Duh, Jennifer K Sun, and Alan W Stitt. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. *JCI Insight*, 2(14), July 2017.

[Fed21] International Diabetes Federation. Idf diabetes atlas, 2021.

[FFAH22] Mohamed M Farag, Mariam Fouad, and Amr T Abdel-Hamid. Automatic severity classification of diabetic retinopathy based on densenet and convolutional block attention module. *IEEE Access*, 10, 2022.

[GL17] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, July 2017.

[GPC$^+$16] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), December 2016.

[HGL$^+$23] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, February 2023.

[HJS22] Abid Haleem, Mohd Javaid, and Ravi Pratap Singh. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(4), October 2022.

[HSCA16] Mark B Horton, Paolo S Silva, Jerry D Cavallerano, and Lloyd Paul Aiello. Clinical components of telemedicine programs for diabetic retinopathy. *Current Diabetes Reports*, 16(12):129, December 2016.

[HST$^+$22] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), April 2022.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXive*, 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXive*, March 2016.

[IEE$^+$23] Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Najat Drawel, Gaith Rjoub, and Witold Pedrycz. A comprehensive survey on applications of transformers for deep learning tasks. *arXive*, June 2023.

[KOM$^+$20] Yusaku Katada, Nobuhiro Ozawa, Kanato Masayoshi, Yoshiko Ofuji, Kazuo Tsubota, and Toshihide Kurihara. Automatic screening for diabetic retinopathy in interracial fundus images using artificial intelligence. *Intelligence-Based Medicine*, 3-4:100024, December 2020.

[MDB23] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *NATO Advanced Science Institutes Series E: Applied Sciences*, 13(9), April 2023.

[NG18] Mads Fonager Nørgaard and Jakob Grauslund. Automated screening for diabetic retinopathy - a systematic review. *Ophthalmic Research*, 60(1), January 2018.

[NLA$^+$19] Katrine B Nielsen, Mie L Lautrup, Jakob K H Andersen, Thiusius R Savarimuthu, and Jakob Grauslund. Deep Learning-Based algorithms in screening of diabetic retinopathy: A systematic review of diagnostic performance. *Ophthalmology Retina*, 3(4), 2019.

[NPM$^+$22] Dimple Nagpal, S N Panda, Muthukumaran Malarvel, Priyadarshini A Pattanaik, and Mohammad Zubair Khan. A review of diabetic retinopathy: Datasets, approaches, evaluation metrics and future trends. *Journal of King Saud University - Computer and Information Sciences*, 34(9):7138–7152, October 2022.

[NPS22] Onnisa Nanegrungsunk, Direk Patikulsila, and Srinivas R Sadda. Ophthalmic imaging in diabetic retinopathy: A review. *Clin. Experiment. Ophthalmol.*, 50(9), 2022.

[NvGRA07] Meindert Niemeijer, Bram van Ginneken, Maria S A Russell, Stephen Rand Suttorp-Schulten, and Michael D Abràmoff. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Investigative Ophthalmology & Visual Science*, 48(5):2260–2267, May 2007.

[Ope23] OpenAI. GPT-4 technical report. *arXiv*, March 2023.

[PDKB22] Brittney J Palermo, Samantha L D'Amico, Brian Y Kim, and Christopher J Brady. Sensitivity and specificity of handheld fundus cameras for eye disease: A systematic review and pooled analysis. *Surv. Ophthalmol.*, 67(5), 2022.

[RLW$^+$20] Serge Resnikoff, Van Charles Lansingh, Lindsey Washburn, William Felch, Tina-Marie Gauthier, Hugh R Taylor, Kristen Eckert, David Parke, and Peter Wiedemann. Estimated number of ophthalmologists worldwide (international council of ophthalmology update): will we meet the needs? *The British Journal of Ophthalmology*, 104(4):588–592, April 2020.

[RPC+20] Hamza Riaz, Jisu Park, Hojong Choi, Hyunchul Kim, and Jungsuk Kim. Deep and densely connected networks for classification of diabetic retinopathy. *Diagnostics (Basel)*, 10(1), January 2020.

[SAFL10] Nathan Silberman, Kristy Ahlrich, Rob Fergus, and Subramanian Lakshminarayanan. Case for automated detection of diabetic retinopathy, 2010.

[SCD+17] Sharon D Solomon, Emily Chew, Elia J Duh, Lucia Sobrin, Jennifer K Sun, Brian L VanderBeek, Charles C Wykoff, and Thomas W Gardner. Diabetic retinopathy: A position statement by the american diabetes association. *Diabetes Care*, 40(3):412–418, March 2017.

[SKZ+23] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88, August 2023.

[TCD+20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2020.

[TMS20] Borys Tymchenko, Philip Marchenko, and Dmitry Spodarets. Deep learning approach to diabetic retinopathy detection. *9th International Conference on Pattern Recognition Applications and Methods*, pages 501–509, 01 2020.

[TPM23] Maria Tariq, Vasile Palade, and Yingliang Ma. Transfer learning based classification of diabetic retinopathy on the kaggle EyePACS dataset, 2023.

[TTY+21] Zhen Ling Teo, Yih-Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyungtaek Rim, Ning Cheung, Mukharram M Bikbov, Ya Xing Wang, Yating Tang, Yi Lu, Ian Y Wong, Daniel Shu Wei Ting, Gavin Siew Wei Tan, Jost B Jonas, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11), November 2021.

[TW22] Tien-En Tan and Tien Yin Wong. Diabetic retinopathy: Looking forward to 2030. *Frontiers in Endocrinology*, 13:1077669, 2022.

[UDH+04] D Usher, M Dumskyj, M Himaga, T H Williamson, S Nussey, and J Boyce. Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening. *Diabet. Med.*, 21(1):84–90, January 2004.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXive*, June 2017.

[WHX+21] Jianfang Wu, Ruo Hu, Zhenghong Xiao, Jiaxu Chen, and Jingwei Liu. Vision transformer-based recognition of diabetic retinopathy grade. *Medical Physics*, 48(12), December 2021.

[WLZ18] Shaohua Wan, Yan Liang, and Yin Zhang. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers & Electrical Engineering*, 72:274–282, November 2018.

[WSK+18] Tien Y Wong, Jennifer Sun, Ryo Kawasaki, Paisan Ruamviboonsuk, Neeru Gupta, Van Charles Lansingh, Mauricio Maia, Wanjiku Mathenge, Sunil Moreker, Mahi M K Muqit, Serge Resnikoff, Juan Verdaguer, Peiquan Zhao, Frederick Ferris, Lloyd P Aiello, and Hugh R Taylor. Guidelines on diabetic eye care: The international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*, 125(10), October 2018.

[WWC+22] Andrew M Williams, Jared M Weed, Patrick W Commiskey, Gagan Kalra, and Evan L Waxman. Prevalence of diabetic retinopathy and self-reported barriers to eye care among patients with diabetes in the emergency department: the diabetic retinopathy screening in the emergency department (DRS-ED) study. *BMC Ophthalmology*, 22(1):237, May 2022.

[YRK+12] Joanne W Y Yau, Sophie L Rogers, Ryo Kawasaki, Ecosse L Lamoureux, Jonathan W Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M Dekker, Astrid Fletcher, Jakob Grauslund, Steven Haffner, Richard F Hamman, M Kamran Ikram, Takamasa Kayama, Barbara E K Klein, Ronald Klein, Sannapaneni Krishnaiah, Korapat Mayurasakorn, Joseph P O'Hare, Trevor J Orchard, Massimo Porta, Mohan Rema, Monique S Roy, Tarun Sharma, Jonathan Shaw, Hugh Taylor, James M Tielsch, Rohit Varma, Jie Jin Wang, Ningli Wang, Sheila West, Liang Xu, Miho Yasuda, Xinzhi Zhang, Paul Mitchell, Tien Y Wong, and Meta-Analysis for Eye Disease (META-EYE) Study Group. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, 35(3):556–564, March 2012.

[YXK17] Shuang Yu, Di Xiao, and Yogesan Kanagasingam. Exudate detection for diabetic retinopathy with convolutional neural networks. *2017 39th Annual International Conference of the IEEE*

Engineering in Medicine and Biology Society (EMBC), pages 1744–1747, 2017.

[ZLLS21] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into deep learning. *arXiv*, June 2021.