RESEARCH ARTICLE

Diabetic Retinopathy Detection and Analysis with Convolutional Neural Networks and Vision Transformer

Yogesh Tewari[1], Nitin Singh Parihar[1], Karan Rautela[1], Nishant Kaundal[1], Manoj Diwakar[1,*] and Neeraj Kumar Pandey[1]

[1] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India

Abstract

Diabetic Retinopathy occurs when elevated blood sugar levels damage retinal blood vessels, potentially leading to vision impairment. In this paper, we have tested the performance of CNN, ViT and their hybrid models. The dataset used is publicly available on Kaggle and the dataset contained around 35,000 retinal images which were divided into 5 classes namely No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. In CNN we tested 4 different architectures in which we achieved the best accuracy of 75.4% with Resnet50 architecture and with ViT model we achieved an accuracy of 83.9% and from the hybrid model we achieved an accuracy of 88.4% from the Resnet50 + ViT. The results shown by the models were promising but there were some gaps in the study. The dataset used was skewed towards NO DR class. For future work more balanced datasets with some data augmentation techniques could be used. Additionally, the study used only 50 epochs which can be increased in future work to use the model to their full potential.

Keywords: diabetic retinopathy, CNN, ViT, deep learning, image classification.

1 Introduction

Diabetic retinopathy (DR) is the condition which happens when there is a sudden increase in the blood sugar level which drastically has a great impact on the blood vessel of retina. It significantly affects individuals with diabetes, potentially causing complete vision loss and blindness. Early detection during the initial stages (Mild DR) can prevent progression to severe vision loss. Starting from mild DR to moderate DR as the blood sugar level increases, the growth of tissue also increases. The final stage is Proliferative DR where growth of such abnormal tissue increases leading to complete vision loss. Thus, detection of such disease in early stages can be an impactful way to reduce diabetic retinopathy. However, there are many areas where the doctors who treat eyes particularly called ophthalmologists and proper equipment are not available. The growth in population of people who are drastically being affected by diabetic retinopathy is rising which eventually leads to a shortage of healthcare centers. Eventually the eye doctor's time will devote to how to cure it rather than classifying it. Thus, making a work for

Submitted: 30 March 2025
Accepted: 07 May 2025
Published: 03 June 2025

Vol. 1, **No.** 1, 2025.
10.62762/BISH.2025.724307

***Corresponding author:**
✉ Manoj Diwakar
manoj.diwakar@gmail.com

Citation
Tewari, Y., Parihar, N. S., Rautela, K., Kaundal, N., Diwakar, M., & Pandey, N. K. (2025). Diabetic Retinopathy Detection and Analysis with Convolutional Neural Networks and Vision Transformer. *Biomedical Informatics and Smart Healthcare*, 1(1), 18–26.

AI tools to solve this classifying problem. There are various neural networks which provide efficient results in the realm of classification problems. We can now easily, with the help of computer vision tools, automate the process, reducing the burden on doctors. It is now possible to classify different stages of DR, which will help in early diagnosis and give further recommendations. In recent years we have seen progress in deep learning for medical image analysis, especially for retinal disease diagnosis. In this study we compare and analyze different deep learning models including Convolutional Neural Network (CNN), Vision Transformers (ViT), a hybrid model of ResNet-50+ViT and lastly hybrid model of EfficientNet-B0+ViT combination. These models are evaluated on a Kaggle dataset naming Diabetic Retinopathy (resized) which consist of almost 35000 retinal images labelled according to five diabetic retinopathy stages. Convolutional Neural Networks (CNNs) are widely used for image classification tasks due to their high feature extraction capabilities. They consist of different layers, namely convolutional layers used for detection of patterns like edge, features, structure of retinal images. Reduction of spatial dimension while keeping essential features is done by pooling layer. Fully connected layer used for classifying the extracted feature to one of the five diabetic retinopathy stages. CNN are helpful in capturing local spatial features but make struggle in long range dependency in a complex image. Vision Transformers (ViTs) have significantly gained popularity as an alternative of CNN as it offers an alternative approach by utilizing self-attention mechanism to support long range dependency across a complex retinal image. Unlike CNN, which processes an image using spatial hierarchies, ViT divides the retinal image into patches of fixed size which are thereby used as tokens in a sequence. Then these patches are carried forward through a series of multi head self-attention layers which helps in understanding relationship between different parts of retinal image. ViTs are helpful where capture of contextual relationships is crucial. However, it takes a large amount of training data and computational power to give results. To overcome the limitation of independent CNN and ViT, we implement a hybrid model that integrates both architectures. One such model is a hybrid of ResNet-50 and ViT. ResNet-50, a deep CNN architecture with residual connection, used for feature extraction and learning capabilities. By combination of this hybrid model uses CNN's local feature extraction while utilizing ViT's global self-attention mechanisms for better image classification. Combination of these two helps in capturing diabetic retinopathy related abnormalities which could not be done by CNN independently. Additionally, we implemented another hybrid model that combines EfficientNet-B0 with ViT. EfficientNet-B0 is an optimization of CNN architecture which balances accuracy and computational efficiency by scaling depth, width and resolution in a structured manner. By this combination it reduces computational costs thus making it promising for diabetic retinopathy analysis. This study aims to make a comparison of these four-model architecture-CNN, ViT, ResNet-50+ViT and EfficientNet-B0+ViT in detecting and classifying on a publicly available Kaggle dataset consisting of retinal images labelled for five classes. Each model is evaluated to determine which efficient approach for diabetic retinopathy detection. By developing such a reliable diabetic retinopathy detection deep learning system, this research helps ophthalmologists for early detection and management of diabetic retinopathy and contribute to AI-driven advancement on retinal images.

2 Related Work

In the area of Healthcare, Abramoff et al. [1] explored a key trail on Autonomous AI-Based Diagnostic System for Diabetic Retinopathy. It helps people to get primary care without specialist consultations. This research highlighted the system's high accuracy and efficiency in Determining DR. It also shows the importance of integrating such AI systems into the healthcare sector for early detection and management of DR. The study conducted by Ting et al. [2] reviews the global prevalence of DR. It also reviewed the risk factors associated. It shows the lack of facilities across the country for screening practices. They show the need for a robust public health program to address DR detection and treatment obstacle. In this competition conducted by Kaggle named APTOS-2019 Challenge Kartik [3] developed AI model for DR Detection .In this report , he was provided by dataset with DR severity levels ranging from 0(no DR) to 4(Proliferative DR).On basis of various performance matrices like accuracy , sensitivity and integrity he established a robust dataset and evaluation framework which helps researchers to advance DR detection models.

The research Enhanced U-Net for Diabetic Retinopathy Segmentation conducted by Agarwal [4] showcased

the DR detection by performing Image segmentation on Retinal images using enhanced U-Net model. He used IDRiD dataset to perform lesion segmentation. By using this he was able achieve higher accuracy compared to traditional U-Net models as it isolated affected regions in retinal images increases better preprocessing. Dosovitskiy et al. [5] proposed the Vision Transformer (ViT) in their work An Image is Worth 16×16 Words: Transformers for Image Recognition. The model applies a self-attention mechanism to image recognition tasks, enabling it to effectively capture both global and local features. Their study demonstrated the scalability and superior performance of Transformer-based architectures compared to conventional convolutional neural networks (CNNs). Gulshan et al. [6] used the Harris Hawk Optimization (HHO) algorithm which was inspired by hawks' cooperative hunting strategies. In this they applied HHO to hyperparameter tuning which helps in performance of neural network. By this study, they able to demonstrate potential for optimizing complex models to increase performance like ViT in DR detection tasks. Zhai et al. [7] discussed various methods to enhance Vision Transformers for tackling larger datasets. They improved stability and computational efficiency for large scale image recognition tasks. By doing this ViTs make them ideal for DR detection where larger datasets like APTOS-2019 are commonly in use.

In the study published by Kobat et al. [8] shows detection of DR using Pre-Trained DenseNet with Digital Fundus Images. Here they use horizontal and vertical patch division. This model extracts deep features in both three class and five class classifications. The patching and hybrid model increases localization and robustness. Tanlikesmath et al. [9] created the dataset of eye images for doing study for diabetic retinopathy analysis. He used both resized and cropped images for dataset creation. By this he was able to create a data set consisting of 35 thousand images. Vaswani et al. [10] in field of transformer model, they marked a go off from RNNs and CNNs. In their study they set up benchmarks for NLP tasks. Their efforts continue to refine the transformer and explore in areas such as computer vision. It helps to extend impacts in language processing.

In the study published by Li et al. [11] they use a dataset of 13673 images from 9598 patients. Additionally, 757 images were manually annotated for lesion detection. By using this dataset, they achieve an accuracy of 0.8284.Despite high classification accuracy. model failed with precise lesion localization. It underscored the complexity of the task. Under this study Staal et al. [12] they performed the retinal vessel segmentation for early detection of diabetic retinopathy. To solve this problem, they used a ridge-based segmentation approach. They used KNN classifier for the following approach combined with sequential forward feature selection. Performance was evaluated on a dataset of 40 manually labelled retinal images. It achieved a ROC curve of 0.952 and accuracy of 0.944. Quellec et al. [13] used deep image mining for diabetic retinopathy screening. They trained ConvNet to detect image with DR. In this study they performed supervised with image refer to DR or not only. They used a public dataset of 90000 images. This model is at last able to outperform lesion detectors in DiaretDB1 dataset. Szegedy et al. [14] they stated that CNN have revolutionized the image recognition, that enables in advancement of accuracy and efficiency. By this study, they concluded that Inception network optimizes computational efficiency with carrying high accuracy, whereas ResNet allows for deeper models with improved gradient flow. This study leads to work on hybrid models in future to enhance the result and accuracy. Decencière et al. [15] created a hybrid model BrownViTNet to solve the problem of Brownfield site detection. They created an architecture consisting of four initial convolutional layers with intermediate layers using a Vision transformer (ViT).AS a result, this model achieves faster convergence and better generalization as compared to simple CNN models. It also improved feature representations. Sugeno et al. [16] used publicly available Kaggle Asia Pacific Tele-Ophthalmology Society APTOS 2019 training dataset. They used EfficientNet-B3 model and train model on this dataset. They achieved classification accuracy for top predicated label as 0.84,0.95 for second prediction and 0.98 for third prediction. To enhance DR analysis, they performed lesion detection. Key observations under this are simultaneous detection of blood vessels and red lesions, accurate extraction of white lesions and validation using DIARETDB1 dataset. Usman et al. [17] applied three differ states of the art CNN architectures, name as ResNet50, ResNet152 and SqueezeNet1 to classify the lesions. Under this study ResNet50 achieved accuracy of 93.67%, SqueezeNet1 achieved accuracy of 91.94% and ResNet152 achieved highest accuracy among all three of 94.40%.

In the study achieved by Willis et al. [18] reinforces the critical role of early detection and severity assessment

in helping the patients. In this study they analysis 1004 adults aged 40 and above with diabetics. Dr Severity marked using Early Treatment Diabetic Retinopathy Study (ETDRS) severity scale. It helps them to refining deep learning model for DR classification, developing treatment strategies. The studies reviewed in this survey show improvement of DR detection using various models such as shown in [19–24]. It helps in enhancing technology by using diverse datasets. It helps to establish applications and detect them without specialist requirements.

3 Methodology

Chronic diabetes elevates blood sugar levels, damaging retinal blood vessels. This can affect the retina by either blurriness or in some cases complete vision loss. This research provides the necessity of Convolutional Neural Networks (CNNs) and Vision Transformer (ViTs) in detecting the impact of diabetes on retina. The essential steps here used are data collection, data preprocessing, model training and then comparing all of them together and choosing which one of the following will be best for our study.

3.1 Data Collection

The dataset that we have used here is downloaded from Kaggle named Diabetic Retinopathy (Resized) which consists of 35126 unique retinal images which were labelled on five classes namely No DR (0), Mild DR (1), Moderate DR (2), Severe DR (3), Proliferative DR (4). The size of the dataset is 7.95GB. Below are some attached retinal images from the dataset.

3.2 Data Pre-Processing

In this step we did data augmentation which basically means flipping, rescaling, adding some angle and finally resizing the image into size of 224 X 224 pixels. The data is split into the ratio of 60:20:20 which means approximate 21000 images were kept for data training whereas for data testing we have 7000 images and for the validation part we have 7000 retinal images. These models CNN and ViT cannot process the image directly, so we need to convert them. For CNN we convert the image into NumPy array of shape (224,224,3) and labelled them into 5 classes by one hot encoding. On the other hand, ViT accepts the input in the form of tensor (3,224,224) and labels the images in form of integers 0 for No DR, 1 for Mild DR, 2 for Moderate DR, 3 for Severe DR and 4 for Proliferative DR. Figure 1 shows the sample retinal images.

3.3 Model Architecture

3.3.1 ResNet-50

ResNet-50 as shown in Figure 2, is a deep convolutional neural network (CNN) that aims to address the vanishing gradient issue in deep networks through residual learning. ResNet-50 consists of 50 layers, including convolutional layers, batch normalization layers, activation functions, and skip connections (identity shortcuts). Efficient gradient flow is facilitated through the skip connections, and therefore deeper networks can be trained without compromising performance. The model receives input images resized to $224 \times 224 \times 3$ that first pass through an initial 7×7-sized convolutional layer, batch normalization, and then a ReLU activation function. A max-pooling layer subsequently compresses spatial dimensions before the data are fed into main residual blocks. ResNet-50 consists of four stages each containing multiple residual blocks. A block contains three convolutional layers which are a 1×1 convolution for dimension reduction, a 3×3 convolution for feature extraction and a 1×1 convolution to restore dimensions. At the end of the network, a global average pooling layer, followed by a fully connected layer and a softmax activation function, is employed for diabetic retinopathy classification. Due to its capability of extracting hierarchical image features, ResNet-50 can successfully detect both global retinal structures and local lesions and is hence appropriate for retinal image analysis.

3.3.2 Vision Transformer (ViT)

The Vision Transformer (ViT) model as shown in Figure 3, used in this study is a transformer-based model for image classification that effectively captures global and local contextual information. In contrast to standard Convolutional Neural Networks (CNNs) that use convolutional layers, ViT processes images as a sequence of non-overlapping patches and uses self-attention mechanisms to learn feature representations. The input image resized to size $224 \times 224 \times 3$ is divided into 16×16 patches, and 196 patches are tiled in a grid of size 14×14. The patch is flattened and projected into a 768-dimensional embedding using a linear projection layer. For preserving spatial relations, learnable positional embeddings are incorporated with patch embeddings. In addition, another classification token (CLS token) is appended at the end of the input sequence, which summarizes information from all the patches and goes through self-attention. The most critical part of the model is multiple transformer encoder layers, each

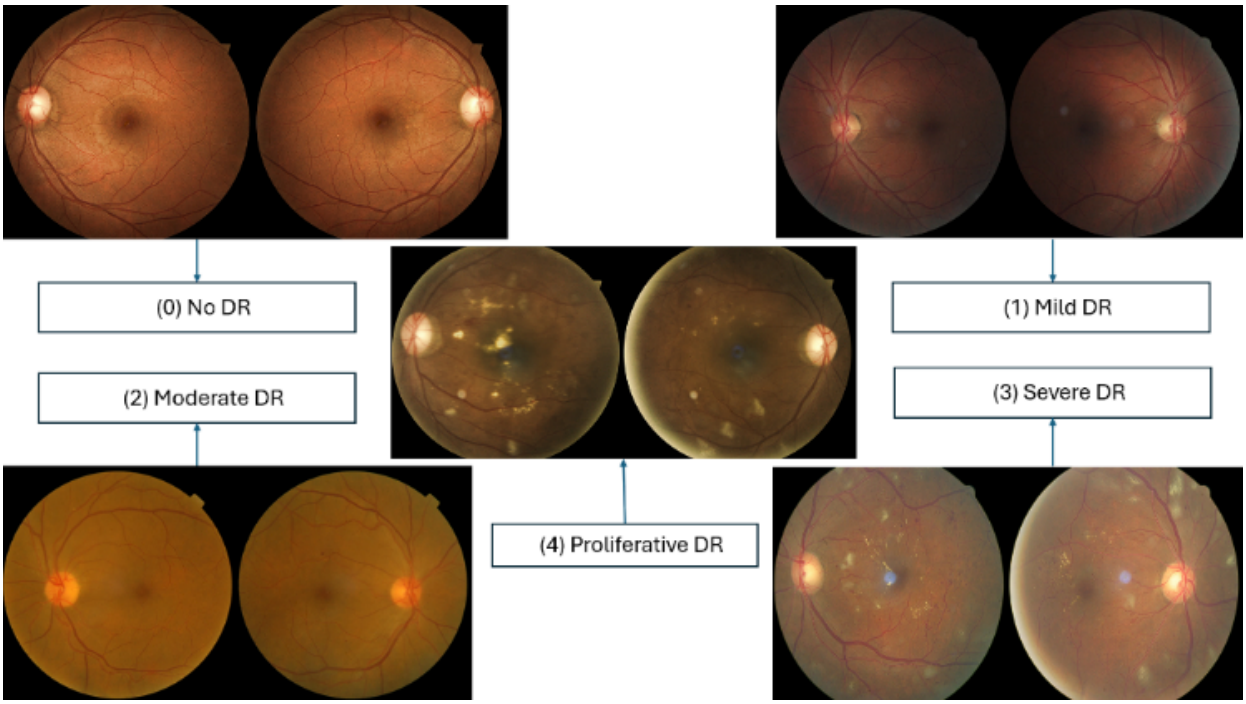

Figure 1. Retinal Images from dataset.

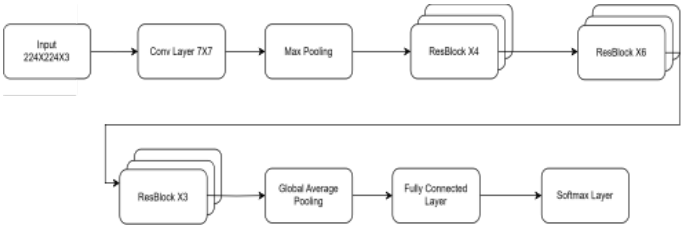

Figure 2. ResNet-50 Architecture.

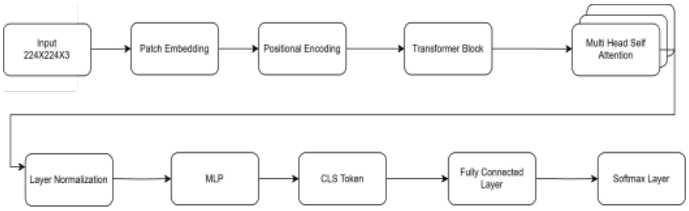

Figure 3. ViT Architecture.

enhance classification accuracy. Its ability to replicate world feature dependencies supports its efficiency in detecting subtle retinal pathologies, such that it is particularly superior to CNNs, especially in the context of a vast dataset.

3.3.3 ResNet-50 + Vision Transformer (ViT) Hybrid Model

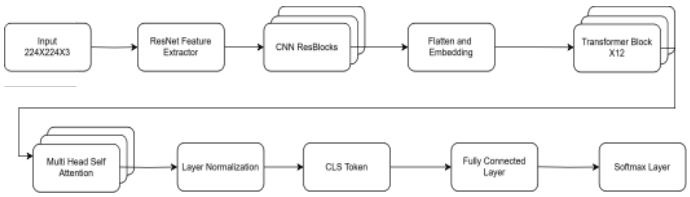

Figure 4. ResNet-50 + ViT Architecture.

containing Multi-Head Self-Attention (MHSA) and Feed-Forward Neural Networks (FFN). The MHSA process enables the model to capture long-range dependencies between image regions, and hence the model is particularly effective in diabetic retinopathy feature detection. Layer Normalization (LN) applies to each transformer block for input normalization and applies Dropout for overfitting prevention. Following the transformer layers, the CLS token output is subjected to a fully connected layer and softmax activation function to classify the image into one of five grades of diabetic retinopathy severity. The ViT model is pre-trained on large datasets before fine-tuning on the diabetic retinopathy dataset to

The ResNet-50 + ViT model as shown in Figure 4, leverages the strengths of Convolutional Neural Networks (CNNs) in local feature extraction and Vision Transformers (ViTs) for global contextual information capture. The model has two main components:

CNN Feature Extractor: Convolutional backbone, i.e., ResNet-50, converts the input image to capture local features such as microaneurysms, hemorrhages, and exudates. The ReLU activations and batch normalization with max-pooling are utilized in the convolutional layers to down-sample the spatial dimensions. The feature maps resulting from CNN are flattened and reshaped into sequential patches to

be used as input for the transformer module.

ViT Transformer Encoder: The ViT block consumes these feature patches obtained from CNN as a sequence and utilizes Multi-Head Self-Attention (MHSA) for learning the long-range dependencies. A classification token (CLS token) and position embeddings help learn inter-relation across different retinal areas. The final classification layer consists of a fully connected layer and softmax activation function for predicting diabetic retinopathy severity.

This combination strategy effectively integrates CNNs to learn fine-grained spatial features and the self-attention mechanism of ViT to learn global dependencies, resulting in improved classification accuracy.

3.3.4 *EfficientNet-B0 + ViT Hybrid Model*

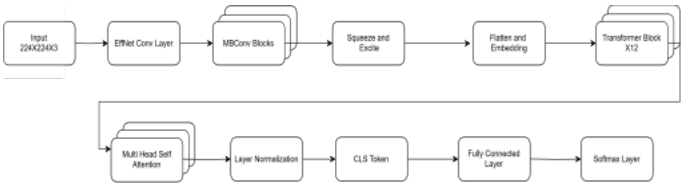

Figure 5. EfficientNet-B0 + ViT Architecture.

EfficientNet-B0 + ViT model as shown in Figure 5, incorporates EfficientNet-B0 and a Vision Transformer (ViT) to generate a lightweight but efficient framework for detecting diabetic retinopathy. EfficientNet-B0 utilizes compound scaling to balance depth, width, and resolution in achieving maximum accuracy with the aim of ensuring computational efficiency. It begins with a 3×3 convolutional stem, 3×3 batch normalization, and Swish activation, and employs MBConv blocks with squeeze-and-excitation (SE) mechanisms to effectively extract retinal abnormalities. The extracted features are sequentially flattened into patches and fed into the ViT transformer encoder, which executes Multi-Head Self-Attention (MHSA), feed-forward layers, and layer normalization to capture global dependencies. A classification token (CLS token) gathers significant information, which is then passed on to a softmax activated fully connected layer to predict the image belonging to one out of five grades of diabetic retinopathy severity. With their excellent local feature extraction and global contextual learning, this hybrid model synergistically enhances classification accuracy with low computational costs.

3.4 Training Process

Training the ResNet-50, ViT, ResNet-50 + ViT, and EfficientNet-B0 + ViT models includes preprocessing data, augmentation, optimization, and evaluation. First, images are resized to 224×224, normalized between $[0, 1]$, randomly flipped, rotated $(\pm 20^{\circ})$, zoomed at $(0.8x - -1.2x)$, contrast modified, and added with Gaussian noise. The dataset is split into 70% training, 15% validation, and 15% testing. The models are optimized using the Adam optimizer (learning rate $= 0.0001$, decayed by cosine annealing) with categorical cross-entropy loss and class weighting. Training is performed with a batch size of 64 for 50 epochs with dropout (0.5), L2 weight decay $(1e - 4)$, batch normalization (for CNNs), and layer normalization (for ViTs). Gradient clipping (max norm $= 1$) is performed in ViTs to prevent exploding gradients. During the forward pass, CNNs extract local hierarchical features and ViTs employ self-attention to learn global dependencies. The output logits are passed through softmax, and the loss is backpropagated using automatic differentiation. We track training and validation loss, as well as training and validation accuracy, and use early stopping if there is no improvement after 10 epochs. Learning Rate Scheduler reduces the learning rate by 0.1 each time it finds a plateau, to achieve stable convergence.

4 Results

This section presents the results of the four models—ResNet-50, Vision Transformer (ViT), ResNet-50 + ViT, and EfficientNetB0 + ViT—on diabetic retinopathy classification. The models were trained on the dataset, and their performance was evaluated as shown in Table 1, based on key metrics such as accuracy, loss, precision, recall, and validation performance.

4.1 ResNet-50

ResNet-50 model is used for training datasets to evaluate their performance parameters. By using this model, we achieved training accuracy of 98.72%, training loss of 0.0363, validation accuracy of 75.40% and it gives validation loss of 1.6320.These results are shown in the form of graph as shown in Figure 6.

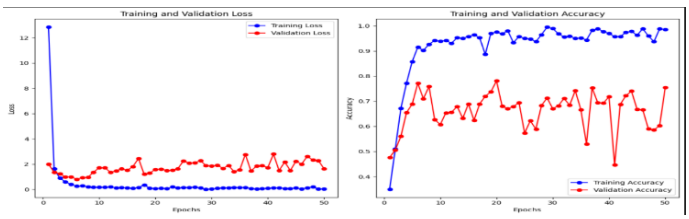

Figure 6. ResNet-50.

Table 1. Comparison of all four models.

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
ResNet-50	98.72%	0.0363	75.40%	1.6320
ViT	99.93%	0.0030	83.90%	1.2849
ResNet-50 + ViT	99.26%	0.0235	88.40%	0.8906
EfficientNet-B0+ViT	98.91%	0.0331	87.00%	0.8766

4.2 Vision Transformer (ViT)

ViT model is used for training datasets to evaluate their performance parameters. By using this model, we achieved training accuracy of 99.93%, training loss of 0.0030, validation accuracy of 83.90% and it gives validation loss of 1.2849.These results are shown in the form of graph as shown in Figure 7.

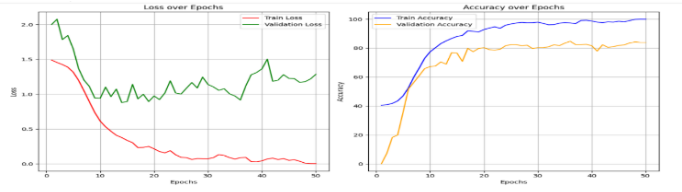

Figure 7. ViT.

4.3 ResNet-50 + Vision Transformer (ViT) Hybrid Model

The Hybrid model of ResNet-50 and ViT is used for training datasets to evaluate their performance parameters. By using this model, we achieved training accuracy of 99.26%, training loss of 0.0235, validation accuracy of 88.40% and it gives validation loss of 0.8906.These results are shown in the form of the graph as shown in Figure 8.

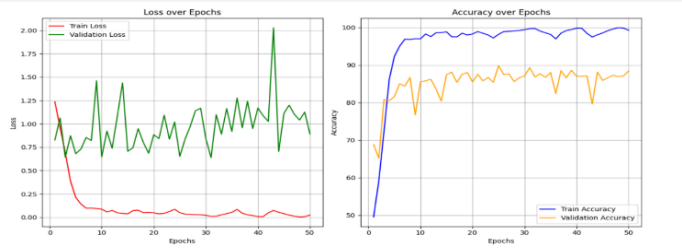

Figure 8. ResNet-50 + ViT.

4.4 EfficientNet-B0 + ViT Hybrid Model

The Hybrid model of EfficientNetB0 and ViT is used for training datasets to evaluate their performance parameters. By using this model, we achieved training accuracy of 98.91%, training loss of 0.0331, validation accuracy of 87.00% and it gives validation loss of 0.8766.These results are shown in the form of graph as shown in Figure 9.

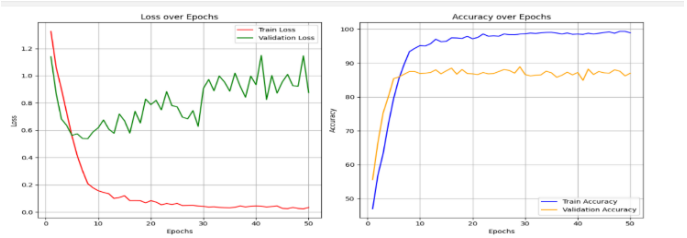

Figure 9. EfficientNet-B0 + ViT.

5 Conclusion

The results of this study demonstrate that hybrid models combining CNNs and Vision Transformers outperform standalone models in diabetic retinopathy classification, which is a complex medical image classification task. Among the four models tested, the ResNet-50 + ViT hybrid emerged as the best-performing model, achieving the highest validation accuracy (88.40%) and the lowest validation loss (0.8906). This model's ability to leverage both local feature extraction and global context understanding enables it to capture subtle and complex patterns in retinal images, making it the most effective for detecting diabetic retinopathy.

The EfficientNetB0 + ViT hybrid model, while slightly less accurate, still provided excellent performance with a validation accuracy of 87.00%. The advantage of this model lies in its efficiency—its reduced computational cost makes it ideal for resource-constrained environments, such as mobile or edge computing devices, without sacrificing too much in terms of classification accuracy.

In comparison, the standalone CNN model (ResNet-50) struggled with generalization, achieving lower validation accuracy (75.40%) and higher validation loss (1.6320), while the standalone ViT model showed better performance (83.90% validation accuracy), demonstrating the importance of global context learning in this task.

Overall, this study confirms that combining CNNs with Vision Transformers provides a robust and efficient solution for diabetic retinopathy classification,

leveraging the strengths of both architectures. These hybrid models represent a promising direction for future research and practical deployment in medical image analysis, particularly in diagnosing diabetic retinopathy, where both local abnormalities and global contextual information are critical for accurate detection. Future work could explore additional optimizations, such as hybridizing other state-of-the-art models or incorporating attention mechanisms that are specifically designed for medical image tasks.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine, 1*(1), 39. [Crossref]

[2] Ting, D. S. W., Cheung, G. C. M., & Wong, T. Y. (2016). Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & experimental ophthalmology, 44*(4), 260-277. [Crossref]

[3] S.D. Karthik Maggie. APTOS 2019 Blindness Detection. *Kaggle* (2019).

[4] Agarwal, R. (2023, November). Diabetic retinopathy segmentation in IDRiD using enhanced U-Net. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)* (pp. 1-6). IEEE. [Crossref]

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [Crossref]

[6] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama, 316*(22), 2402-2410. [Crossref]

[7] Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12104-12113). [Crossref]

[8] Kobat, S. G., Baygin, N., Yusufoglu, E., Baygin, M., Barua, P. D., Dogan, S., ... & Acharya, U. R. (2022). Automated diabetic retinopathy detection using horizontal and vertical patch division-based pre-trained DenseNET with digital fundus images. *Diagnostics, 12*(8), 1975. [Crossref]

[9] Tanlikesmath. Diabetic Retinopathy Detection Competition Dataset Resized/Cropped (2019).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*. [Crossref]

[11] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences, 501*, 511-522. [Crossref]

[12] Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., & Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging, 23*(4), 501-509. [Crossref]

[13] Quellec, G., Charriere, K., Boudi, Y., Cochener, B., & Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. *Medical image analysis, 39*, 178-193. [Crossref]

[14] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1). [Crossref]

[15] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., ... & Klein, J. C. (2014). Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology*, 231-234. [Crossref]

[16] Sugeno, A., Ishikawa, Y., Ohshima, T., & Muramatsu, R. (2021). Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in biology and medicine, 137*, 104795. [Crossref]

[17] Usman, T. M., Saheed, Y. K., Ignace, D., & Nsang, A. (2023). Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal of Cognitive Computing in Engineering, 4*, 78-88. [Crossref]

[18] Willis, J. R., Doan, Q. V., Gleeson, M., Haskova, Z., Ramulu, P., Morse, L., & Cantrell, R. A. (2017). Vision-related functional burden of diabetic retinopathy across severity levels in the United States. *JAMA ophthalmology, 135*(9), 926-932.

[Crossref]

[19] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, *3*(3), 25. [Crossref]

[20] Wu, Y., Xia, Y., Song, Y., Zhang, Y., & Cai, W. (2020). NFN+: A novel network followed network for retinal vessel segmentation. *Neural Networks*, *126*, 153-162. [Crossref]

[21] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141). [Crossref]

[22] Song, J., Zheng, Y., Wang, J., Zakir Ullah, M., & Jiao, W. (2021). Multicolor image classification using the multimodal information bottleneck network (MMIB-Net) for detecting diabetic retinopathy. *Optics Express*, *29*(14), 22732-22748. [Crossref]

[23] Li, X., Hu, X., Yu, L., Zhu, L., Fu, C. W., & Heng, P. A. (2019). CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, *39*(5), 1483-1493. [Crossref]

[24] Mo, J., Zhang, L., & Feng, Y. (2018). Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks. *Neurocomputing*, *290*, 161-171. [Crossref]

Yogesh Tewari is currently student of Computer Science and Engineering Department in Graphic Era University, Dehradun, India. (Email: yogeshtewari370@gmail.com)

Nitin Singh Parihar is currently student of Computer Science and Engineering Department in Graphic Era University Dehradun, India. (Email: pariharnitin033@gmail.com)

Karan Rautela is currently student of Computer Science and Engineering Department in Graphic Era University Dehradun, India. (E-mail: karanrautela.097@gmail.com)

Nishant Kaundal is currently student of Computer Science and Engineering Department in Graphic Era University Dehradun, India. (Email: nishantkaundal18@gmail.com)

Dr. Manoj Diwakar is a Professor in the Department of Computer Science and Engineering at Graphic Era Deemed to be University, Dehradun, India. With over 16 years of academic and industrial experience, he has significantly contributed to the fields of image processing, computer vision, medical imaging, and information security. Throughout his career, Dr. Diwakar has published over 200 research papers in esteemed journals and conferences, including those by IEEE, Elsevier, and Springer. He has also served as an editor and reviewer for several reputed journals and has organized numerous international conferences. (Email: manoj.diwakar@gmail.com)

Dr. Neeraj Kumar Pandey is an Associate Professor in the Department of Computer Science & Engineering at Graphic Era (Deemed to be) University, Dehradun, India. Dr. Pandey actively contributes to the academic community as a reviewer and editorial board member for several international journals. He has also served on organizing and review committees for numerous IEEE conferences. His dedication to education and research was recognized when he received the "Best Academic Coordinator" award at the 4th International Annual Awards Summit, Global Education and Corporate Leadership (GECL-2021). (Email: dr.neerajkpandey@gmail.com)