Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 03, September 2025.

A HYBRID CNN-TRANSFORMER APPROACH FOR PRECISE THREE-CLASS DIABETIC RETINOPATHY CLASSIFICATION

Samira Ait Kaci Azzou[1], Djamila Boukredera[2] and Sifeddine Baouz[3]

(Received: 4-Feb.-2025, Revised: 11-Apr.-2025, Accepted: 16-Apr.-2025)

ABSTRACT

This study evaluates the effectiveness of Vision Transformers (ViTs) and hybrid deep-learning architectures for diabetic retinopathy (DR) classification, addressing the challenge of inter-stage ambiguity in traditional systems. While convolutional neural networks (CNNs) such as ResNet50 excel at localized feature extraction in retinal images, ViTs offer superior global contextual modeling. To synergize these strengths, we propose a hybrid architecture integrating ResNet50's granular feature extraction with ViTs' global relational reasoning. Three models are designed and evaluated: (1) an auto-tuned ResNet50, (2) a hyperparameter-optimized ViT and (3) a hybrid model combining both architectures. To reduce ambiguity between neighboring stages, we simplified the traditional five-stage classification into three clinically relevant categories: no DR, early DR (mild/moderate) and advanced DR (severe/proliferative). Trained and validated on the APTOS dataset, the ResNet50 model achieves precision scores of 93.0% (No DR), 82.0% (Early DR) and 86.0% (Advanced DR). The standalone ViT demonstrates relative improvements, attaining 98.0%, 91.0% and 93.0%, respectively. The hybrid model surpasses both, achieving 98.0% average precision across all classes, with gains of +7.0% (early DR) and +5.0% (advanced DR) over the standalone ViT. The proposed hybrid model achieved an impressive value of 99.5% on all metrics (accuracy, precision and recall) for identifying DR (binary classification) and a value of 98.3% for 3-stage classification. It was also concluded that the proposed method achieved better performance in DR detection and classification compared to conventional CNN and other state-of-the-art methods. The proposed hybrid approach significantly reduces confusion between classes, demonstrating its potential for accurate classification of the different stages of DR.

KEYWORDS

Diabetic retinopathy, Vision transformer, Transfer learning, Artificial intelligence.

1. INTRODUCTION

Diabetic retinopathy (DR) is a disease that affects the blood vessels of the retina and can result in blindness. It is a serious complication in diabetic patients [1]-[2]. DDR is identified by the emergence of several types of lesions on the retina. The lesions include microaneurysms (MAs), hemorrhages (HMs) and soft and hard exudates (EXs) [3]. Positive RD is split into several stages. (1) Microaneurysms indicate the mild phase, (2) Moderate stage reveals a stage where blood vessels begin to lose their ability to transport, (3) Severe stage includes blood vessel obstructions and (4) Proliferative stage represent the advanced phases of RD, as shown in Figure 1.

According to the International Diabetes Federation [4], there are around 537 million diabetics, with this figure anticipated to increase to 643 millions by 2030 and 783 millions by 2045. Furthermore, most individuals with diabetes remain undiagnosed for DR, because this disease is often asymptomatic until an advanced stage [5]. In order to diagnose and treat DR, regular retinal screening is essential for diabetic patients. Classification issues associated with DR can be divided into two categories: binary classification and five-class classification. Binary classification focuses on distinguishing between sick and healthy retinas in color fundus images, as established by [6]-[8]. Conversely, five-class classification methodologies strive to categorize images into five distinct classes: Class 0- no DR, Class 1- mild DR, Class 2- moderate DR, Class 3- severe DR and Class 4 -proliferative DR [9]-[10], as resumed in Figure 1. Manual examination of retinal images is carried out using traditional methods to detect the presence of DR, which requires experienced and professional ophthalmologists. In

1. S. Ait Kaci Azzou is with University of Bejaia, Faculty of Exact Sciences, LIMED Laboratory, 06000 Bejaia, Algeria. Email: samira.aitkaciazzou@univ-bejaia.dz
2. D. Boukredera is with University of Bejaia, Faculty of Exact Sciences, LMA Laboratory, 06000 Bejaia, Algeria. Email: djamila.boukredera@univ-bejaia.dz
3. S. Baouz is with University of Bejaia, Faculty of Exact Sciences, Department of Computer Science, 06000 Bejaia, Algeria. Email: sifeddine.baouz@se.univ-bejaia.dz

addition, there is a high probability of misdiagnosis during the manual examination, which is time-consuming and costly.

Automated methods have emerged as viable solutions to enable early identification of Diabetic Retinopathy (DR) and avoid permanent blindness [11]-[12], overcoming problems related to manual classification. In this case, machine learning has shown to be the most effective technique to overcome this problem [13].

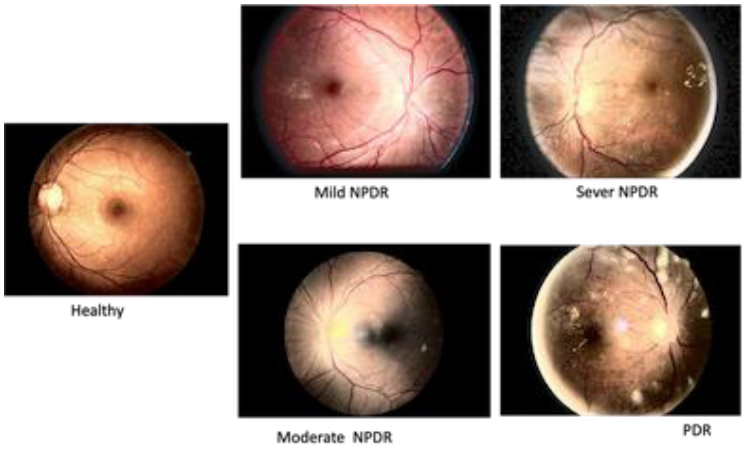


Figure 1. Fundus images representing phases of diabetic retinopathy from the Aptos dataset.

Deep-learning (DL) methods, particularly transfer-learning models, like VGG16, InceptionV3 and ResNet50, have shown considerable promise in analyzing medical images [14-[17]. Convolutional neural networks (CNNs), which underpin these models, mainly concentrate on local features in the input images, which restricts their capability to effectively recognize long-range dependencies and global contextual connections. Vision Transformers (ViTs) have emerged as a revolutionary substitute, addressing these constraints by utilizing self-attention mechanisms to capture long-range dependencies and global contextual associations throughout whole images. While transfer learning-based approaches [18]-[19] have been widely adopted for diabetic-retinopathy (DR) severity classification, existing methods struggle with diagnostic accuracy in early-stage DR, where subtle lesion patterns (e.g. microaneurysms, mild hemorrhages) necessitate both fine-grained feature extraction and global contextual understanding of the retinal image.

To address these challenges and evaluate the effectiveness of ViTs for DR classification, we propose and compare three architectures, each differing in its feature extraction method:

1) ResNet50-based model: A CNN baseline optimized *via* Bayesian hyperparameter tuning for localized feature extraction.
2) ViT-based model:A standalone ViT model tailored for global dependency modeling.
3) Hybrid architecture: A novel fusion of ResNet50 and ViT, combining their complementary strengths.

We further redefine the traditional five-stage DR grading system into three clinically relevant classes: no DR, early DR (encompassing mild and moderate stages) and advanced DR (comprising severe and proliferative stages). This regrouping minimizes confusion between closely related stages, enhancing classification accuracy. Experiments carried out on the APTOS 2019 dataset [20] demonstrate that the hybrid architecture achieves 98.0% precision across all classes, reducing misclassification between adjacent stages by 15%–20% compared to standalone models. ViTs alone outperform ResNet50, with relative improvements of 11.0% (early DR) and 8.1% (advanced DR) in precision. The hybrid architecture significantly enhances early-stage detection of DR, leading to better clinical results.

To sum up, our contributions are as follows:

1) Three novel architectures for DR detection and classification:
   - AtRD/AtR3C: Auto-tuned ResNet50 models with Bayesian hyper-parameter optimization, achieving 99.22% detection accuracy and 94.26% 3-class severity-classification accuracy.
   - ViRD/ViR3C: Vision Transformer (ViT) models leveraging global attention, attaining 97.73%

detection accuracy and 92.97% classification accuracy.
- Revi-RD/Revi-3C: A hybrid CNN-Transformer architecture combining both precedent architectures. It achieves 99.55% detection accuracy and 98.26% 3-class severity-classification accuracy.

2) Redefined DR grading into (0: no DR, 1: early DR, 2: advanced DR), reducing ambiguity in traditional 5-stage grading between neighbor classes.

3) State-of-the-Art Performance:
   - The proposed models are validated on the APTOS 2019 dataset and compared against one another, highlighting the effectiveness of ViTs and the complementary advantages of the CNN-ViT hybrid architecture.
   - The earliest stages were detected with greater accuracy, especially in the hybrid model.
   - We effectively optimized each model's performance as compared to previous methodologies. With the hybrid approach, we greatly outperformed previous results.

The rest of the article is organized as follows: Section 2 reviews relevant research conducted on the DR classification. Section 3 details the methodology, including data pre-processing, the proposed approach and performance measures. The results are presented and analyzed in Section 4. Finally, Section 5 presents the key conclusions and recommendations for future works.

2. LITERATURE REVIEW

Early identification of Diabetic Retinopathy (DR) remains a significant challenge. Researchers have investigated several techniques to address this issue. Classifying DR from retinal images falls into two main categories. Binary classification determines whether or not DR exists, whereas multi-class classification indicates the disease's specific stage. This latter method needs the model to differentiate minor visual variations between DR stages, making it a more difficult task. Several studies have investigated both binary and 5-class classification of DR using machine-learning (ML) [13], [21]-[22], deep-learning (DL) [14], [23]-[25], transfer-learning techniques (TL) [8], [26]-[28] and more recently vision-transformer methods [29]-[30]. However, research into the classification of DR into three classes remains limited. Public retinal-image datasets, such as Idrid, EyePACS, Messidor and Aptos, have been instrumental in these studies for detecting and diagnosing DR. This work will specifically focus on recent advancements in transfer learning (TL) and Vision Transformer (ViT) applied to DR detection and classification on the Aptos dataset.

2.1 Transfer Learning in DR Classification

Dekhil et al. [31] proposed a customized CNN based on a transfer-learning technique for a 3-class classification task. It consists of a pre-processing stage, VGG16 and fully connected layers. To adapt the pre-trained model, they retrained all the layers, achieving a validation accuracy of 77%. In their study [32], Rao et al. evaluated five CNN classifiers; namely, Inception-V3, VGG19, VGG16, Resnet50 and InceptionResNetV2. Resnet50 achieved the highest accuracy (95.59%) for a binary classification. InceptionResNetV2 excelled at multi-class classification. It reached an accuracy of 88.14% for classifying DR into three stages and 85% accuracy for a five-stage classification. Gangwar and Rav [33] proposed an hybrid model incorporating a custom convolutional neural network (CNN) block added to the pre-trained Inception-ResNet-v2. For training these hybrid models, they utilized two Kaggle datasets: Messidor-1 and the APTOS 2019. The achieved test accuracy was 72.33% for Messidor-1 and 82.18% for the APTOS 2019 dataset, respectively. Islam et al. [34] proposed an architecture based on supervised contrastive learning, utilizing the pre-trained Xception model, the APTOS dataset and Messidor-2. They achieved an accuracy of 98.36% for binary classification and 84.364% for multi-class classification. Their study revealed an improvement in performance compared to previous architectures, including ResNet50, Inception and other earlier models. Oulhadj et al. [35] proposed an automatic method based on deep learning. It consists of two main steps; the first one is the pre-processing. The second one is the classification. Four CNN models (Densenet-121, Xception, Inception-v3 and Resnet50) are employed to detect the DR-severity stage. The authors implemented a voting mechanism using the APTOS 2019 dataset. They achieved a final accuracy of 85.28%. Mondal et al. [36] also suggested a deep-learning strategy for detecting diabetic retinopathy that combines the DenseNet101 and ResNet models. Experiments were carried out using the APTOS19 and

DIARETDB1 datasets. Their approach produced an accuracy of 86.08% for five-class classification and 96.98% for binary classification. Many CNN-based techniques have proved their ability to extract subtle image features surpassing traditional methods. While CNNs excel at extracting discriminative local features, crucial for recognizing subtle image characteristics, they struggle to process long-range information due to their inherent local receptive field mechanism. This limitation hinders their ability to fully understand the complex patterns associated with diabetic retinopathy. To address CNNs' difficulties in collecting long-range dependencies within retinal images, Vision Transformers (ViTs) have emerged as a potential solution.

2.2 ViT in DR Classification

Dosovitskiy et al. [37] introduced the Vision Transformer (ViT) for image classification, motivated by the effectiveness of transformers in natural-language processing [38]. ViTs have surpassed traditional convolutional neural networks in a variety of computer-vision tasks by considering images as sequences of patches and exploiting self-attention. Despite the promising potential of ViTs, their application in DR classification remains relatively unexplored and studies specifically focused on DR classification are still limited. Recently, the remarkable representation capabilities of transformers received increasing interest in medical-image analysis [39]-[40]. For DR classification, Wu et al. [41] employed ViTs to prove their superior performance compared to CNNs. Additionally, Mohan et al. [42] proved that dividing the fundus images into non-overlapping portions maintains information about the position of each patch. A different dataset was used to test the effectiveness of DR classification. For example, Nazih et al. [43] provided a ViT-based deep-learning pipeline for recognizing the severity stages of DR. ViT requires big datasets for successful learning; therefore, they utilized the FGADR (fine-grained annotated diabetic retinopathy) dataset, which comprises 1,842 fundus images, to build their model. Experimental results of their ViT model using F1-score, accuracy, and recall metrics were 82.5%, 82.5% and 82.5%, respectively. In [29], Gu et al. classified DR using ViT on the DDR dataset. The performances of the model using specificity, sensitivity and accuracy metrics were 82.45%, 81.40% and 82.35%, respectively. Khan et al. [44] presented an automated approach for DR-severity classification using a fine-tuned Compact Convolutional Transformer (CCT) model, which combines convolutional layers with transformer mechanisms. The model was trained on a huge dataset created by combining five datasets (Aptos, Idrid, Messidor2, DDR and Kaagel Dr dataset). Different pre-processing and augmentation techniques were used to improve image quality. The model achieved an accuracy of 84.5%, outperforming both the ViT (81.56%) and the shifted window transformer (Swin) (82.23%). Different ViT architectures are tested in the study conducted by Karkera et al. [45]. Four pre-trained image transformers: ViT, DeiT, CaiT and BEiT, were trained on a dataset called DBtr. The researchers then combined all four models to predict the severity stages of DR. The combined approach achieved an accuracy of 94.63% outperforming the results obtained by each of the individual models. Recently, Oulhadj et al. [46], proposed a hybrid architecture combining a fine-tuning vision transformer and a capsule network for automatic prediction of the severity level of diabetic retinopathy. The approach was evaluated using four datasets, including APTOS, Messidor-2, DDR and EyePACS and attained the best accuracy scores on the Aptos dataset: 88.18%. Lian and Liu in [47] combined a convolutional neural network (Inception-Resnet-v2) with a vision transformer. The model attained an accuracy of 93.2% using Messidor1 for binary classification and an accuracy of 89.1% using the Aptos dataset for 5-stage classification. Yang et al. [48] have developed a Transformer model based on multiple instance learning (MIL) to classify diabetic retinopathy (DR). Their model divides high-resolution retinal pictures into 224 × 224 pixel patches, which are then processed by a Vision Transformer (ViT) to extract local characteristics. A Global Instance Computing Block (GICB) then combines information from many patches, improving the model's capacity to understand global relationships within the image. The model obtained 93.2% accuracy for binary classification on the Messidor1 dataset and 85.65% accuracy for 5- stage classification on the Aptos dataset, surpassing the Mil-ViT proposed by Yu et al. [49]. Dihin et al. [50] used a combination of Wavelet and multi-Wavelet transforms with the Swin-transformer model. The study highlights the innovative use of the multi-Wavelet transform for feature extraction, integrated into the Swin transformer. The model obtained 96% accuracy for binary classification on the Kaggle APTOS 2019 dataset. The Swin-T model with multi-Wavelet transformation achieved a 98% recall and 96% F1-score for binary classification. However, the model's accuracy decreased in multi-class classification (82%). Approaches based solely on CNNs or ViTs struggle to combine the detection of local lesions

with the analysis of the global anatomical context, which accentuates the ambiguity between classes. To demonstrate the efficacy of hybridization in overcoming these limitations, this study proposes a hybrid CNN-ViT architecture that combines fine feature extraction and contextual modeling. Further, we redefine DR staging into three-tier clinically actionable categories - no DR, early DR and advanced DR - to improve the accuracy of classification, which remains under-explored in the literature.

3. METHODOLOGY

This section presents three deep-learning architectures for the classification of diabetic retinopathy (DR). Each model was trained for binary detection (0: No DR, 1: DR) and three-stage severity classification (0: No DR, 1: Early DR, 2: Advanced DR). The first proposed architecture employs transfer learning with ResNet-50 for feature extraction. AtRD and AtR3C, respectively, handle binary and 3-class classification. The second proposed architecture uses ViTs for feature extraction. ViRD and ViR3C deal with binary and 3-class classification, respectively. Finally, we propose a hybrid architectures, ReVi-RD and ReVi-3C, for detection and 3-class classification, respectively, combining the strengths of both previous models. As illustrated in Figure 2, each model follows a similar pipeline composed of several processes:

- Pre-processing process that balances the dataset and enhances the quality of input images.
- Feature extraction is performed using the chosen architecturen (Rsnet50 and ViT).
- A multi-layer neural network classifies the image into two or 3-class classification. In the following part, we give more details for each of these processes.

3.1 Datset Description

A Kaggle dataset titled APTOS 2019 Blindness Detection (APTOS stands for Asia Pacific Tele Ophthalmology Society) was used to train and evaluate the models [20]. This dataset was collected by Aravind Eye Hospital in rural areas of India with the objective of developing high-performance tools for the automated diagnosis of diabetic retinopathy and enhancing the hospital's ability to identify potential patients. The dataset consists of 3,662 retinal images, categorized into five stages of diabetic retinopathy (DR)(see Figure 3b): no DR, mild DR, moderate DR, severe DR and proliferative DR, which are annotated with values ranging from 0 to 4. However, one of the main limitations of this dataset is the significant class imbalance, particularly for the severe NPDR category, which contains only 193 images. Additionally, the images vary in size and exhibit considerable variations due to their collection in a real-world multi-center environment. These variations arise from differences in camera settings across centers and the presence of noise, both in the data and in the annotations.

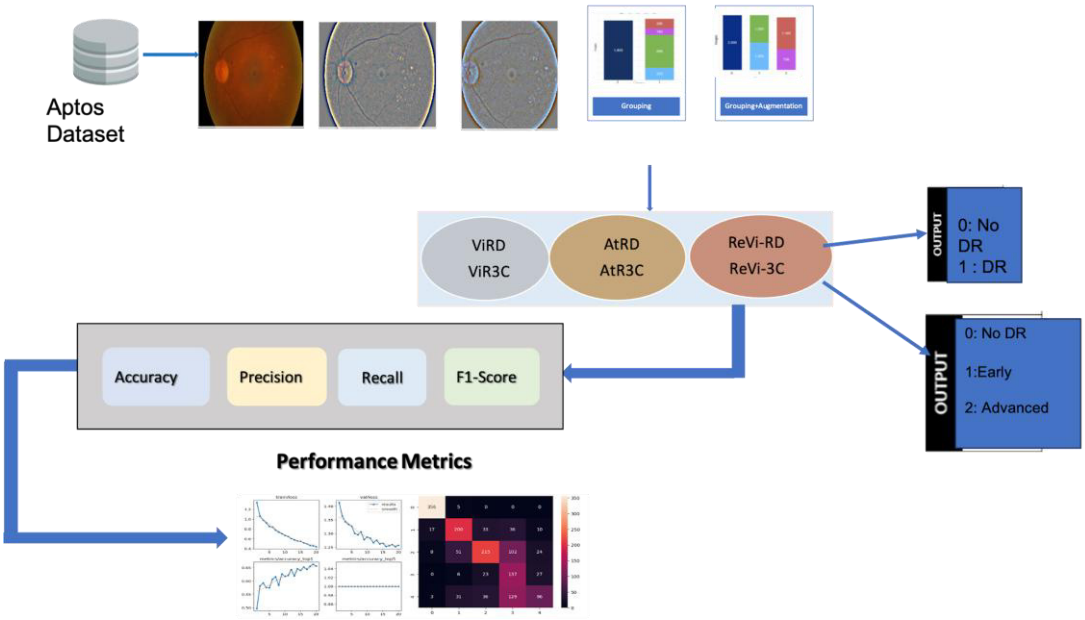
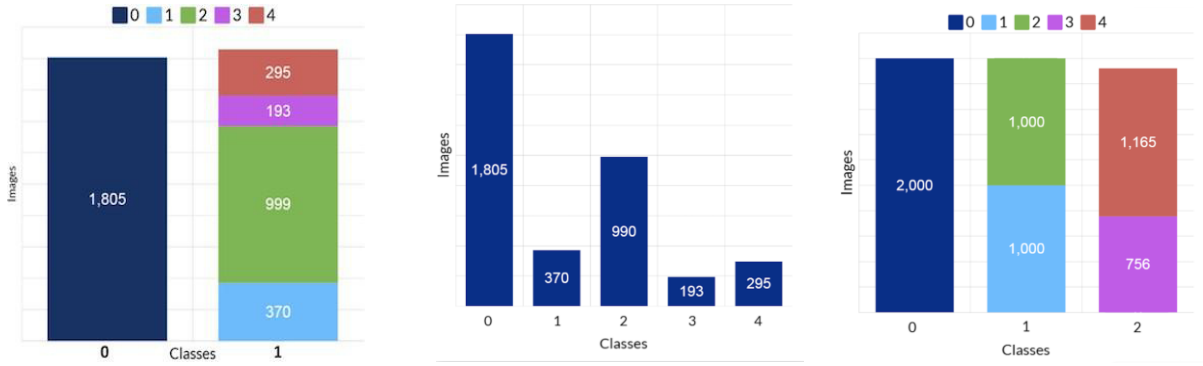


Figure 2. Proposed-approach pipeline from data pre-processing to class prediction.

3.2 Dataset Preparation

Our goal is to develop a model that can detect the existence of DR and classify its severity. As shown in Figure 3a and Figure 3b, the classes were grouped and re-annotated according to the classification task (binary or three-class classification, respectively). However, achieving an accurate model performance necessitates overcoming the persistent problem of data imbalance. For DR detection, we use a binary classification (No DR, DR). This grouping successfully balances the dataset, as shown in Figure 3a.



(a) Binary Aptos dataset (b) Aptos dataset before augmentation (c) 3-class Aptos dataset

Figure 3. Aptos dataset before and after aggregation and augmentation.

However, for three-stage classification, the problem of data imbalance persists. To address this issue, we use data-augmentation techniques that create additional images.

3.3 Data Augmentation

We employ data-augmentation techniques to expand the database and provide additional images of the different DR stages as illustrated in Figure 3c. Each original image underwent multiple augmentation transformations, resulting in five augmented images. These transformations include distortions, horizontal and vertical flips, as well as brightness adjustments. The purpose is twofold: expanding the dataset's variability while meticulously preserving the essential DR characteristics. This enables machine-learning models to learn and identify retinopathy features regardless of the image's position or lighting conditions. Figure 4 shows a sub-set of the generated images by the augmentation process.

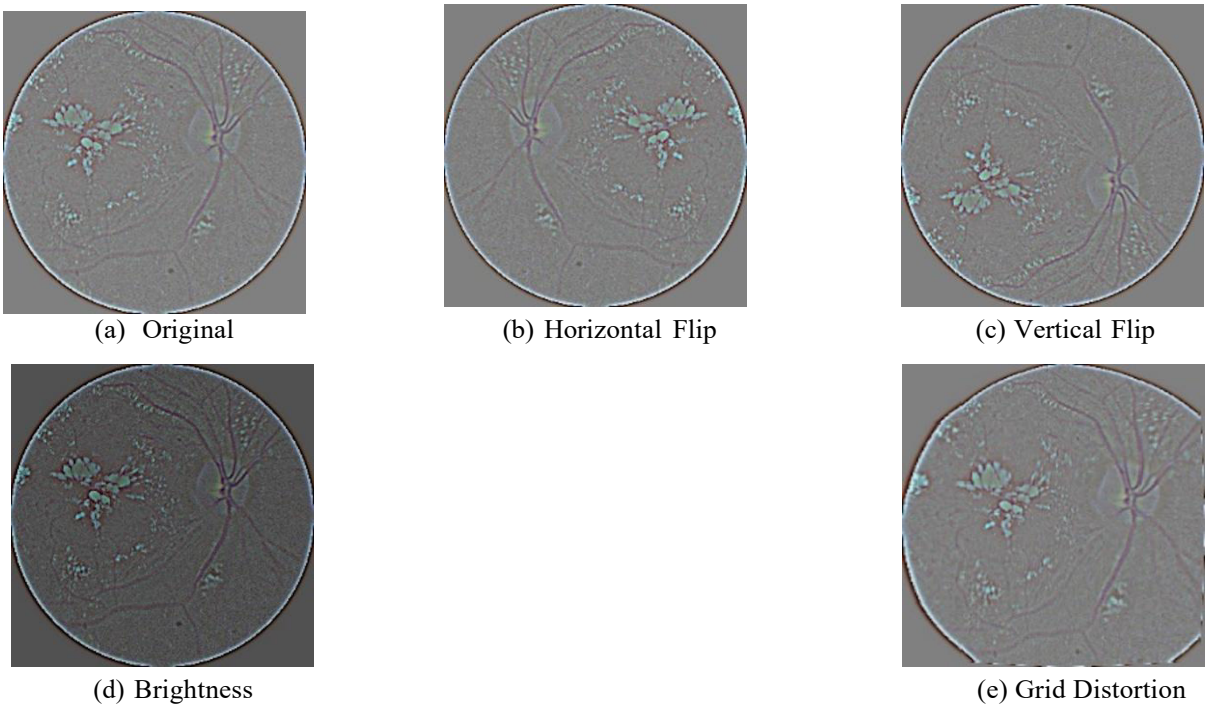
3.4 Image Pre-processing

Due to their many sources, the fundus images in the dataset show significant heterogeneity in terms of size, noise levels and distortion. These variations present significant problems for accurate analysis and reliable lesion detection. To overcome these obstacles and improve the quality of feature extraction, we propose a multi-stage pre-processing process (see Figure 5). The different stages of pre-processing that we have carried out are:

1) The initial step involves resizing all images to a uniform size of 224x224 pixels. This standardization facilitates subsequent analyses and the extraction of characteristics.
2) Each resized color image was converted into gray scale, followed by convolution using a Gaussian blur filter, as illustrated in Figure 5b [51].This step is designed to reduce noise and accentuate features, in particular by improving the visibility of exudate, red lesions and blood vessels.
3) A circular-cropping [52] technique was used to remove non-informative black pixels (background or noise) and retain only the regions of interest, as shown in Figure 5c.
4) Finally, normalization was performed on the pre-processed images to ensure consistent scaling of all pixel values, thereby enhancing the efficiency and stability of model training. This data normalization process aims to standardize the distribution of the images.

3.5 Fine Tuning

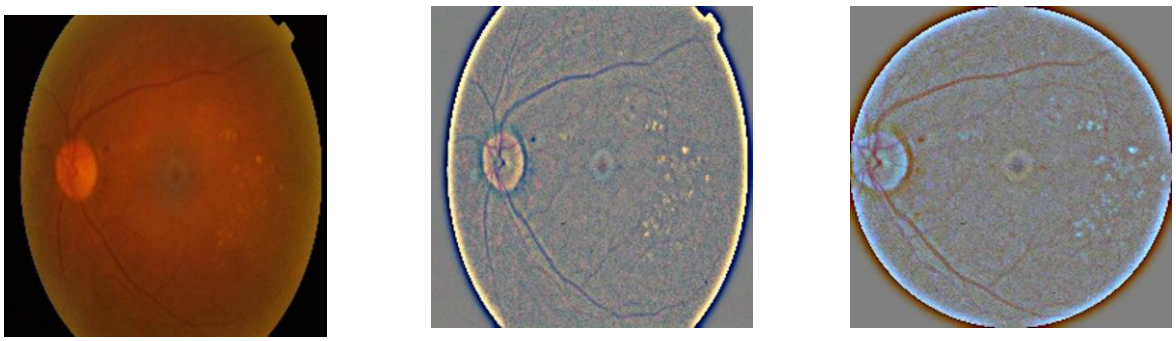
Pre-trained models, such as ResNet50 and Vision Transformers (ViTs), require fine-tuning to meet the specific demands of DR detection and classification. For proposed models—AtRD/AtR3C (ResNet50-



(a) Original (b) Horizontal Flip (c) Vertical Flip

(d) Brightness (e) Grid Distortion

Figure 4. Data-augmentation illustration.



(a) Original (b) Gaussian Blur (c) Circle Crop

Figure 5. Pre-processing phases.

based) and ViRD/ViR3C (ViT-based), we employed a two-phase optimization. First, the pre-trained architectures were fine-tuned on the APTOS dataset, enabling them to capture discriminative retinal features, such as microaneurysms, hemorrhages and exudates, by adapting their weights to the morphological patterns of DR. Second, we applied Bayesian optimization to systematically refine critical hyperparameters, including image resolution, batch size and learning rate, ensuring robust classification performance across DR-severity classes while minimizing overfitting. This dual-phase strategy optimizes both the models' feature-extraction capabilities and training dynamics.

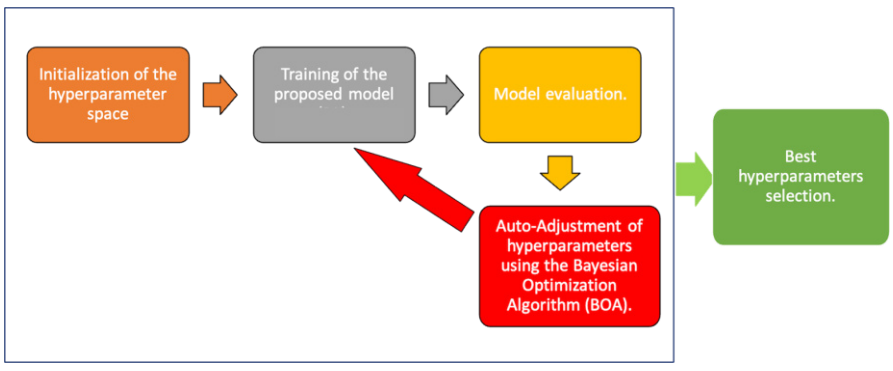


Figure 6. Auto-hyperparameter-tuning process.

As shown in Figure 6, the fine-tuning process using Bayesian optimization aims to efficiently identify the optimal hyperparameter configuration for our architectures based on transfer learning. For efficient optimization, the network is trained with a limited number of epochs while exploring various hyperparameter combinations within a pre-defined range. This approach prioritizes identifying the hyperparameter set that yields the best score on the validation of metric set.

3.6 DR Classification Using AtRD and AtR3C: Approach-based Transfer Learning

Transfer learning, unlike training from scratch, aims to transfer knowledge that has been learned from another data set to a target problem. In this study, we adopted ResNet50, a convolutional neural network pre-trained on the ImageNet dataset, as the backbone for feature extraction.

ResNet-50 is a specific variant of Residual Neural Networks (ResNets), developed by Kaiming He et al. in 2015 [53] for image recognition. It consists of 50 layers structured into convolutional layers and identity blocks. The key innovation of ResNet-50 lies in the use of residual connections, also known as skip connections (see Figure 7), which enable the network to bypass certain layers. This approach facilitates the training of very deep networks by mitigating the vanishing-gradient problem. ResNet-50 adopts an optimized architecture in which each residual block contains three convolutional layers (1×1, 3×3 and 1×1 convolutions) instead of the two used in earlier ResNet variants. The 1×1 convolutions serve to reduce and expand dimensionality, improving computational efficiency, while the 3×3 convolution captures spatial features. Several factors contribute to the model's success: its large receptive fields, which capture more contextual information for each pixel; the separation between localization and classification stages; its computational efficiency at deeper layers; and its effective encoding schemes that rely on low-complexity arithmetic operations.

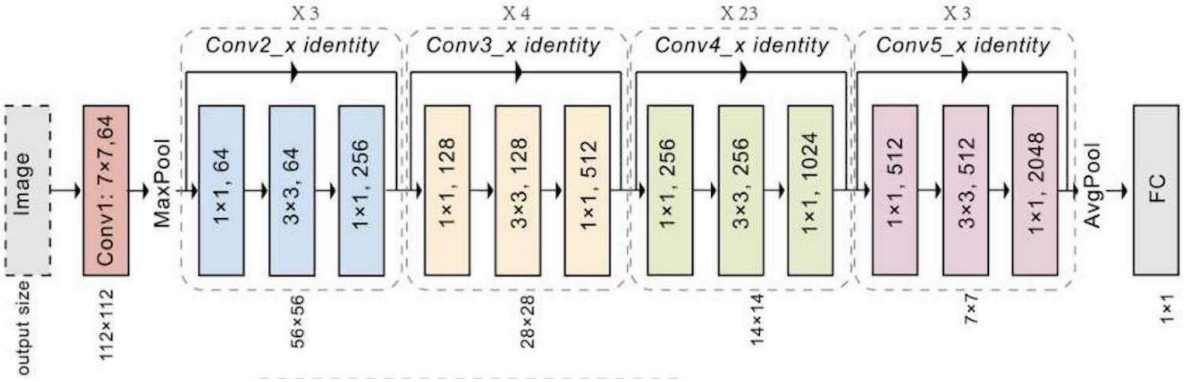


Figure 7. Resnet50 architecture [53].

While ResNet50 excels in general image classification, its final fully connected layer—originally configured for 1,000-class ImageNet classification—was unsuitable for our specialized binary and three-class DR classifications. In response, we designed the AtRD and AtR3C architectures, which retain the feature-extraction capabilities of ResNet50 while incorporating domain-specific adaptations. As illustrated in Figure 8, we replaced the final classification layer of ResNet50 with a customized multi-layer perceptron (MLP) comprising five additional layers (Flatten, Dense, Dropout, Dense, Dense). The final dense layer contains two nodes for binary classification or three nodes for 3-class classification.

3.7 DR Classification Using ViRD and ViR3C: Approach-based ViT

Taking advantage of ViT's ability to model long-range dependencies, we propose ViRD and ViR3C, two ViT-based architectures, for the detection and classification of DR. Figures 9 illustrates the proposed architecture.

The important components of the transformer are multi-head self-attention (MSA) and multi-layer perception (MLP). Multi-head attention in the Figure 10 is the core part of the Transformer. The ViT model considers an image submitted as a series of image patches.

Here are the key steps in its operation:

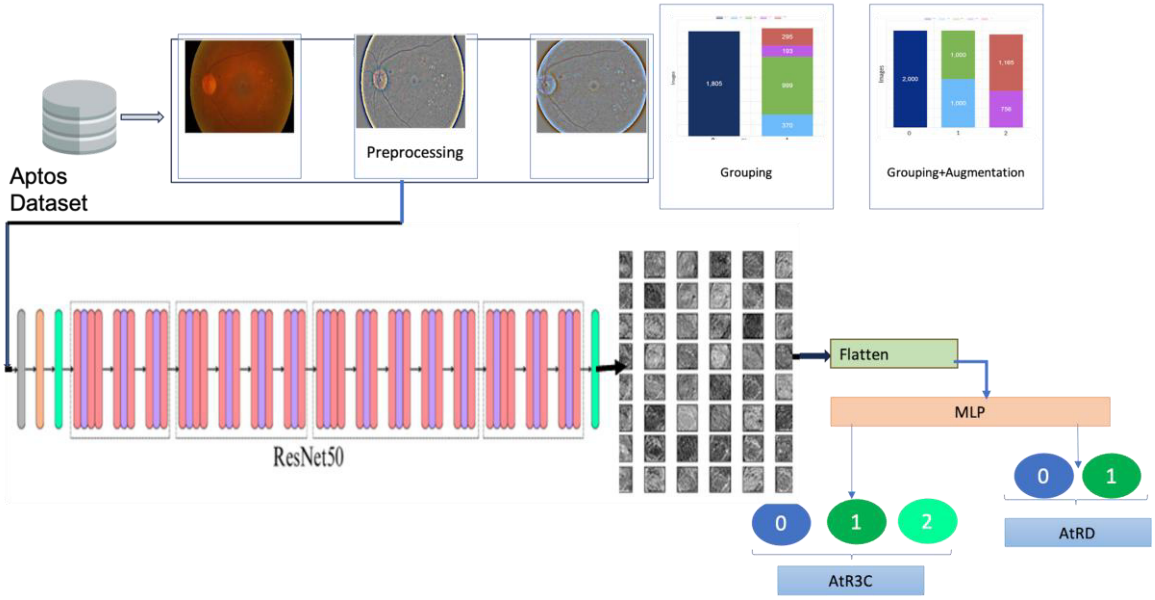


Figure 8. Proposed architecture-based ResNet50: AtRD and AtR3C.

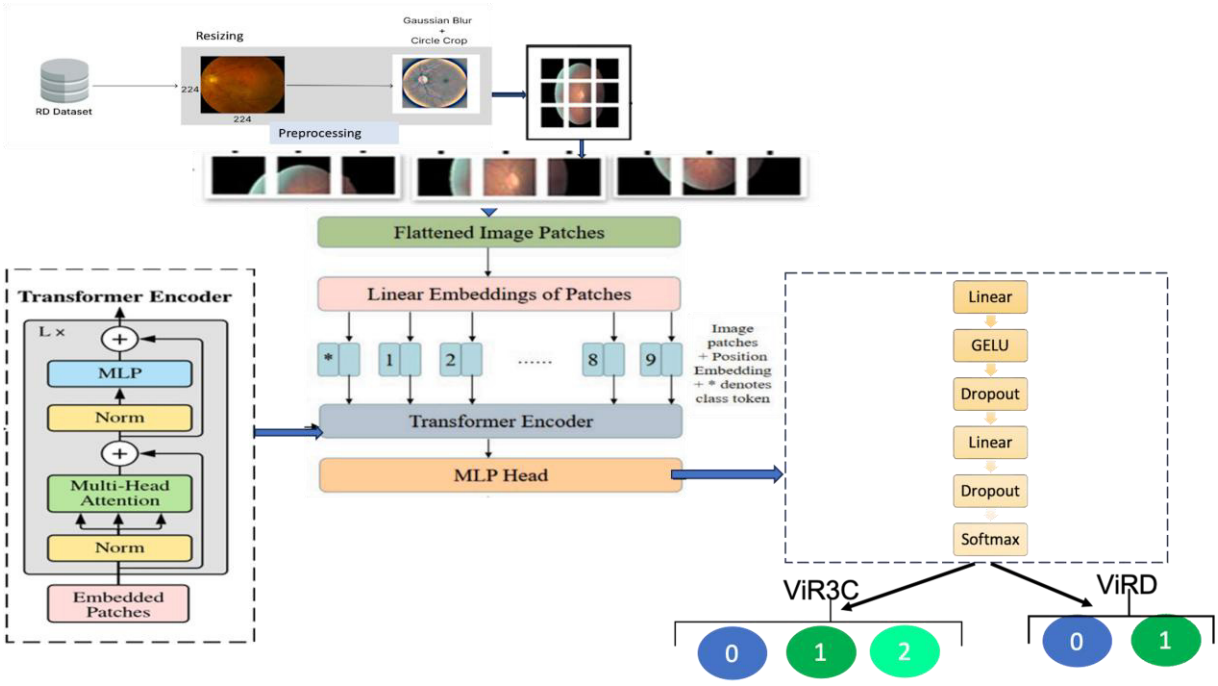


Figure 9. Proposed architecture-based ViT: ViRD and ViR3C.

Image Splitting into Patches: After pre-processing and resizing to 224*224, input picture I is divided into a series of flattened patches Xip (for i = 1, 2, ..., np), each with a size of p × p × C, C=3 corresponding to the three RGB channels in the image I; p = 16, resulting in np =(224 × 224/16*16)= 196 patches. Each patch Xip is flattened and transformed into a 1D vector X_0 of dimension pxpx3= 162x3=768 using linear embedding.

$$X_0 = [x_1, x_2, \dots, x] \in \mathbb{R}^{196\times(768)} \quad (1)$$

Linear Projection of Patches (Patch Embedding): Each flattened patch is projected into a space of dimension *D* using a learnable matrix $E \in \mathbb{R}^{(768)\times D}$. For the *i*-th patch x_i, the embedding is given by $z_i = xi.E$. E represents the projection weight matrix, with dimensions 768×*D*, where 768 is the flattened patch dimension and *D* is the dimension of the projection space. D defines the dimension of the transformer's input tokens, which serve as the basis for self-attention mechanisms. In basic ViTs, D is commonly set to 768.

$$Z_0 = [z_1, z_2, \dots, z_{np}] \in \mathbb{R}^{196\times D} \quad (2)$$

Class Token and Positional Embedding Initialization: As illustrated in Figure 9, the positional information **Pos** $\in \mathbb{R}^{196\times D}$ added into each embedded patch, allowing ViT to better understand the spatial relationships within the input data. The ViT model also incorporates a classification token (z[cls]) inside the embedded patches. This is a randomly initialized, learnable parameter used to aggregate global information for classification. It essentially acts as a decoder.

The input to the Transformer encoder is constructed as:

$$Z = [z[cls], z_0] \in \mathbb{R}^{(196+1\times D)} \tag{3}$$

After adding positional encoding, the final input to the encoder becomes:

$$Z_f = Z + POS \in \mathbb{R}^{(197\times D)} \tag{4}$$

The resulting embedding matrix Z_f, enriched with both visual and positional information, is then fed into a Transformer encoder stack.

Transformer Encoders: The Transformer Encoder is composed of two main layers: Multi-head Self-Attention (MSA) and Multi-layer Perceptron (MLP). The resulting embedding matrix, Z_f , is then fed into a stack of six Transformer encoder blocks. Each block consists of a multi-head self-attention (MSA) module with eight attention heads, followed by a multi-layer perceptron (MLP). Layer normalization and residual connections are applied before and after each sub-layer.

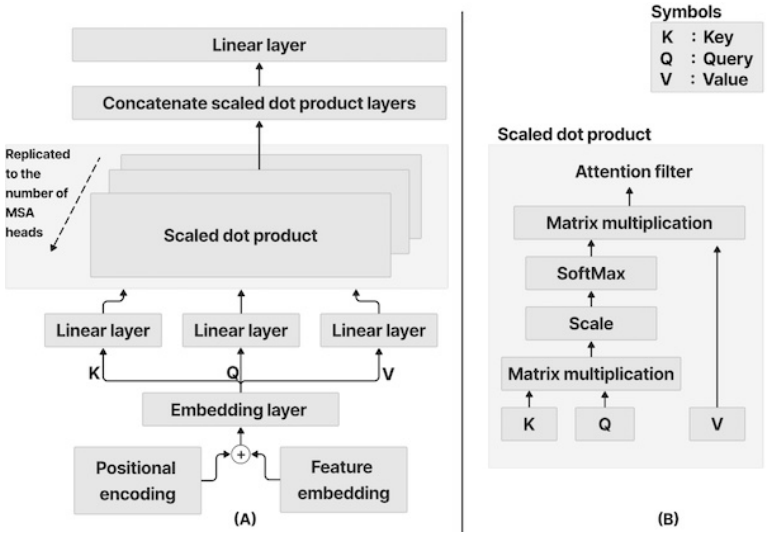


Figure 10. MSA process: (a) MSA process with several attention layers; (b) Scaled dot-product attention [38].

The multi-head attention mechanism (MSA) is a form of self-attention that allows the model to concentrate on information from different sub-spaces of representation at various positions. To calculate attention scores, MSA uses several scaled dot-product attention mechanisms, as shown in Figure 10. The complete MSA operation is summarized as:

$$MSA(Q, K, V) = Concat(h_1, h_2, \dots, h_n).W_0 \tag{5}$$

where *Concat* denotes the concatenation of all attention-head outputs; n is the number of attention heads. h_i is the output of the i-th self-attention head. The concatenated output is then projected back to the original embedding space using a final weight matrix W_0.

The output of each attention head h_i is computed as:

$$h_i = Attention(Q, K, V) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{6}$$

A softmax function is applied to derive the attention weights for the value matrices. This softmax operation normalizes the resulting scores, ensuring that they are positive and sum to unity. We then multiply the attention weights with value matrix (V_i) to get the self-attention output h_i.

The query Q_i, key K_i and value V_i vectors for each head ($i \in \{1, \dots, n\}$) are obtained by multiplying the

input embedding matrix Z_f by three distinct weight matrices, effectively projecting the input embeddings into different representation sub-spaces for each attention head.

$$Q_i = Z_f W_i^Q \qquad K_i = Z_f W_i^K \qquad V_i = Z_f W_i^V$$

The outputs from all the heads are subsequently merged and forwarded to an MLP layer for further processing. Each MLP and MSA operation is preceded and followed by residual blocks and normalization layers to guarantee stability and model optimization. MLP comprises two fully-connected linear layers and between these layers, a non-linear activation function is applied. This function introduces non-linearity, allowing the model to learn more intricate patterns in the data. A common choice for this activation function in ViT is the Gaussian Error Linear Unit (GELU). GELU has a smoother, more continuous shape than the ReLU function, which can make it more effective at learning complex patterns in the data [38].

$$GeLU = 0.5.x + tanh\left[\sqrt{\frac{2}{\pi}}.(x + 0.0447x^2)\right] \quad (7)$$

We introduce two dropout layers to regularize the model and prevent overfitting. Finally, we extract the [Cls] token from the Transformer Encoder output and pass it through a classification head to obtain class predictions y. In order to classify DR into 2 or 3 severity stages, we use a head classification output layer composed of 2 or 3 neurons for ViRD and ViR3C, respectively. We applied a softmax function to get a probability distribution to classify fundus images over the two or three severity stages of DR (see Figure 9).

$$y = softmax(z[Cls]) \quad (8)$$

3.8 DR Classification Using ReVi-RD and ReVi-3C: A Novel Hybrid Approach

To enhance the precision of DR classification, we suggest a novel hybrid architecture that merges the benefits of Vision Transformers (ViTs) and Resnet50. Retinal-image features can be captured locally and globally by ReVi-RD and ReVi-3C models by integrating pre-trained ViRD/ViR3C with pre-trained AtRD/AtR3C models.

The hybrid approach is illustrated in Figure11. To construct this hybrid model, we use the weights of the pre-trained AtRD or AtR3C models to extract local features. We remove the MLP (final layers) of these models and replace it with the pre-trained ViRD or ViR3C, as described in Figure 12. In the following part, we describe our hybrid approach, illustrated in Figure11 and Figure 12, from input images to final classification.

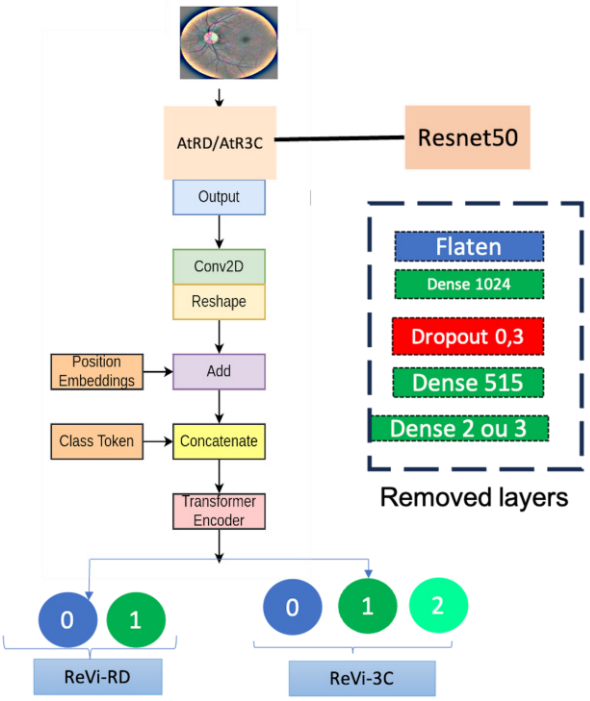


Figure 11. Hybrid architectures: ReVi-RD and ReVi-3C.

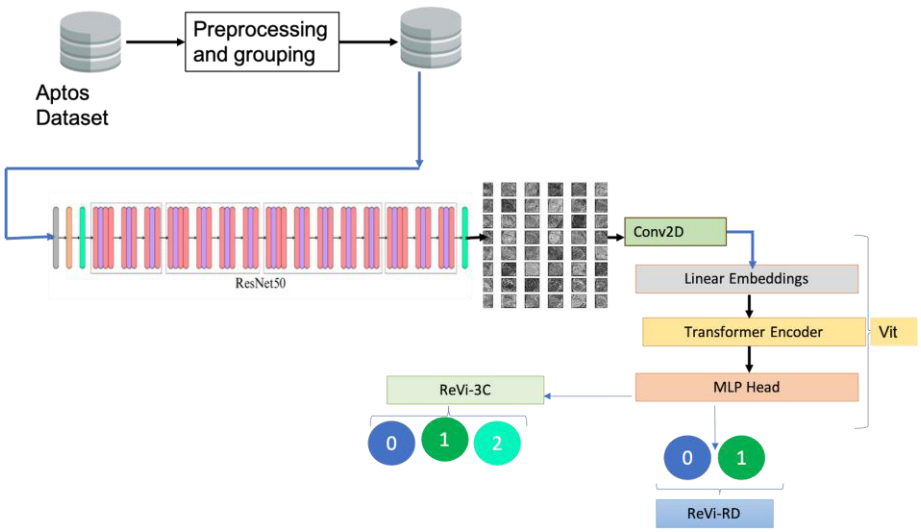


Figure 12. Detailed architecture of ReVi-RD and ReVi-3C.

- Input: an RGB image of 224×224 pixels, represented by a shape tensor [224, 224, 3], which is a standard input size for the ResNet50 model, is introduced into the pre-trained model.
- After pre-processing, AtrD/Atr3C are used to extract local spatial features from input images of size 224×224×3. The final classification layers of AtrD/Atr3C are removed and replaced with a transformer-based head.

  The output from an intermediate layer (specifically, the 7th layer from the end) of the modified ResNet50 model is extracted. The resulting feature map is 7×7×768 in size. This feature map retains high-dimensional representations of localized patterns while compressing spatial resolution to 7×7 grids, each with 768 channels.
- Reshaping for Vision Transformer (ViT): The resulting feature map of dimensions 7×7×768 is reshaped into a sequence of flattened patches, transforming the 7×7×768 feature map into a sequence of 49 tokens, each of 768 dimensions [49, 768]. Here, the 7×7 spatial grid is reinterpreted as 49 non-overlapping "patches", each represented as a 768-dimensional vector. This step adapts the output into a format compatible with transformer-based processing.
- Position Embedding and Class Token: To inject spatial information into the transformer, we add a learnable position embedding to the 49 patches, preserving their spatial relationships. Then, we concatenate a learnable [CLS] token (classification token) to the sequence, increasing its length to 50 ([50, 768]). A final sequence of length 50 is then processed by a Transformer Encoder.
- Transformer Encoder: the sequence of length 50 is fed through a series of 6 Transformer encoder blocks. Each block comprises a multi-head self-attention mechanism with 8 attention heads, followed by an MLP that includes layer normalization and residual connections.
- Classification Head: After the Transformer encoder, we performed a layer normalization and extracted the output corresponding to the class token. Then, we projected the final representation into the class space (2 for ReVi-RD or 3 For ReVi-3C) *via* a dense layer, yielding raw classification scores, which are then transformed into class probabilities using a softmax function.

4. Experimental Results

In this section, a detailed discussion of the experimental results obtained is carried out to prove the effectiveness of the Vits and hybrid models proposed for the classification of DR. The experiment was conducted using the Python environment on a server equipped with an Intel(R) Xeon(R) CPU @ 2.20GHz processor, 13 GB of RAM and a GPU P100 16GB provided by Kaggle platform. We use the Aptos dataset to train and test our architectures. To prevent data leakage, the dataset was explicitly split into two sub-sets with the ratio of 80:20 to make the training and testing datasets. Additionally, to address class imbalance, data augmentation was applied only to the training set, ensuring that artificially generated samples did not leak into validation or test sets.

The model underwent multiple independent trials, each with a unique random seed for dataset shuffling and partitioning. This approach introduced variability in data order and distribution across trials, enabling a thorough assessment of the model's stability.

For the ResNet50-based model, we used the Adam optimizer, while the ViT-based model utilized the AdamW optimizer. We employed categorical cross-entropy as the loss function, suiatable for our multi- class classification task with softmax activation. The learning rate was automatically selected through hyperparameter tuning and the optimal value obtained was 0.0001 for model based on Resnet50 and 0.00002 for model based on ViT. This value was fixed during training to ensure stable convergence.

4.1 Evaluation Metrics

To assess the detection performance of the proposed models, we use the most commonly used metrics: accuracy, precision, specificity or recall (sensitivity) and F1 score. Their mathematical expressions are given in Table 1. TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively.

4.2 Obtained Hyperparameters after Auto-tunning

After image pre-processing, we fine-tuned the architectures to get the best hyperparameters which are presented in Table 2 for AtRD and AtR3C, and in Table 3 for ViRD and ViR3C.

Table 1. Performance metrics.

Metrics	Formula
Accuracy (Acc)	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$
Precision (Positive Predictive Value)	$Precision = \frac{TP}{TP + FP}$
Recall (Sensitivity)	$Recall = \frac{TP}{TP + FN}$
F1-score	$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$
Specificity (True Negative Rate)	$Specificity = \frac{TN}{TN + FP}$

Table 2. Best hyperparameters obtained for AtRD and AtR3C.

Hyperparameter	Value
Image size	224x224
Batch size	32
Warmup epochs	5
Warmup learning rate	0.00001
Epochs	50
Learning rate	0.0001
Weight decay	0.02
Early stopping patience	15
Reduced LR patience	5
Regularizer	0.02

All the proposed architectures are trained using their obtained hyperparameters.

Their performance based on test data was evaluated using the five metrics: accuracy, precision, recall (sensitivity), F1-score and specificity.

4.3 Diabetic Retinopathy Detection Performance

As the first experiment, we compare the performance of AtRD, ViRD and ReVi-RD to evaluate their effectiveness in DR detection and assess the impact of the features extracted by each model. The results reported in Table 4 summarize the evaluation metrics obtained for detecting DR. We can notice that AtRD and ReViRD architectures demonstrate exceptional performance, exceeding 99% across all

metrics (accuracy, precision, recall, F1-score), showcasing their robustness in DR detection. The exceptional performance of AtRD can be attributed to the efficient tuning of hyperparameters. The ViRD model achieves slightly lower, but still impressive results, surpassing 97.7% across all metrics. This disparity arises from the inherent data requirements of ViTs, which typically demand larger datasets to fully leverage their global attention mechanisms compared to transfer-learning models [37]. The hybrid ReViRD model outperforms both standalone architectures , underscoring the synergistic benefits of combining ResNet50's localized feature extraction with ViTs' ability to model long-range dependencies.

The detection performance of the AtRD, ViRD and hybrid ReVi-RD models is compared using their confusion matrices (see Figure 13) and the evaluation metrics summarized in Table 5. The AtRD model achieves high sensitivity in retinopathy detection (99.2% true positive rate), but exhibits a specificity of 96.81%, corresponding to a 3.2% false positive rate in healthy-patient classification. While this underscores its efficacy in identifying pathological cases, the elevated misdiagnosis rate for normal patients highlights limitations in distinguishing subtle non-pathological variations. In contrast, ViRD demonstrates balanced specificity (98.0% overall), with a slightly reduced 2.7% false negative rate for retinopathy cases. Although with an area under curve (AUC) of 99.1% (see Figure 14a), the ViT model is excellent at capturing global context through self-attention; it sometimes misses subtle local features that are critical for identifying retinopathy. This reliance on global context means that, in cases where pathological signs are very localized or subtle, the model might not sufficiently distinguish them from normal variations.

Table 3. Best hyperparameters obtained for the ViRD and Vi3C.

Parameter	Value
Image size	224x224
Batch size	16x16
Train batch size	32
Test batch size	64
Warmup steps	500
Warmup learning rate	0.00001
Epochs	20
Learning rate	0.00002
Weight decay	0.01

Table 4. Performance comparison of proposed models for DR detection (%).

Metric	AtRD	ViRD	ReVi-RD
Accuracy (%)	99.22	97.73	99.55
Precision (%)	99.66	97.72	99.51
Recall (%)	99.23	97.73	99.58
F1-Score (%)	99.40	97.73	99.54
Specificity(Average) (%)	98.01	98.00	99.50

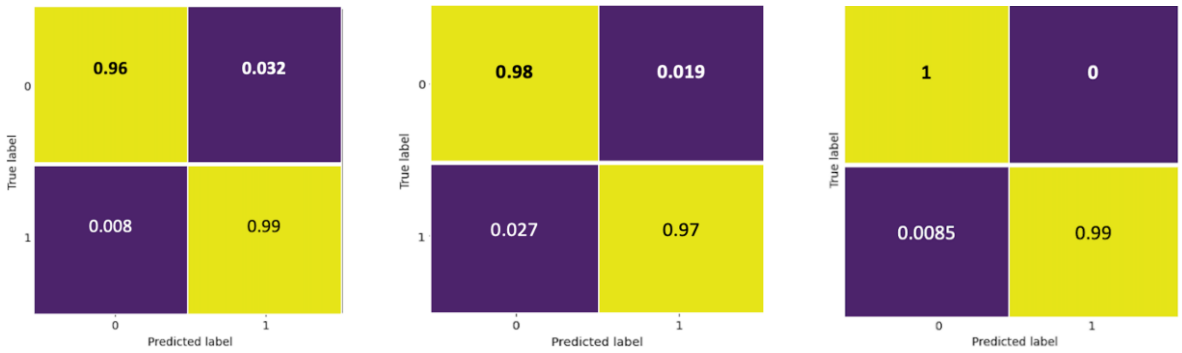
The hybrid ReVi-RD architecture addresses these limitations by synergistically combining CNN-driven local feature extraction (AtRD) and ViT-based global dependency modeling (ViRD). This integration achieves near-perfect classification: a 1.0% false negative rate for retinopathy and 0.0% false positives rate for healthy cases (Table 4). With a specificity of 99.50%, ReVi-RD minimizes unnecessary diagnoses while maintaining exceptional sensitivity, outperforming both AtRD (98.01%) and ViRD (98.00%) in robustness. Class-specific metrics (Table 5) further elucidate these distinctions. AtRD shows moderate precision-recall harmonization (F1-scores: 97.7% for both classes), constrained by CNN architectures' focus on localized textures rather than on lesion correlations. ViRD improves balance, achieving 98.00% F1-scores for both classes *via* global attention, yet remains vulnerable to localized oversights. ReVi-RD's hybrid design transcends these trade-offs, leveraging CNN-localized granularity and ViT-global context to optimize feature representation. This dual capability enables superior accuracy in diabetic-retinopathy classification, particularly for cases requiring simultaneous

fine-grained and global analysis.

The hybrid ReVi-RD resolves residual trade-offs, achieving near-perfect metrics (100% F1-score for both classes, 99–100% precision/recall and 99.9% AUC, as shown in Figure 15a). Its dominance stems from synergizing AtRD localized feature extraction with ViRD global-context modeling, effectively eliminating misclassifications (only 0.85% of non-healthy cases mislabeled). For clinical deployment, ViRD's standalone performance—particularly its precision gains for critical non-healthy cases—validates ViTs as an important tool for severity staging, while ReVi-RD's hybrid architecture sets a new benchmark for applications requiring ultra-reliable classification. These results emphasize the necessity of integrating CNNs and ViTs in medical imaging, where both local granularity and global coherence are essential for accurate, interpretable diagnoses.

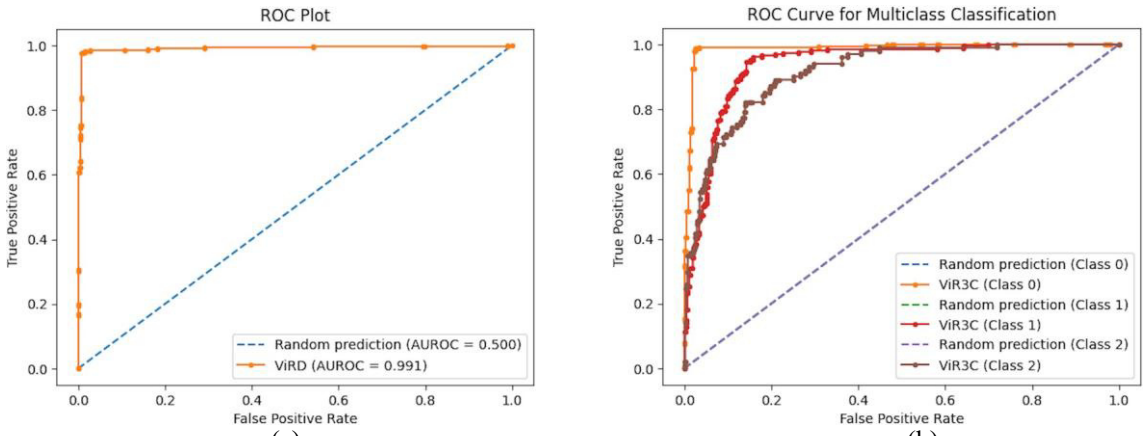
Table 5. Class-wise performance of proposed models for DR detection (%).

	AtRD		ViRD		ReVi-RD	
Metrics	**Class 0**	**Class 1**	**Class 0**	**Class1**	**Class 0**	**Class1**
Precision (%)	97.60	97.90	97.00	98.00	99.00	100.00
Recall (%)	97.90	97.60	98.00	97.00	100.00	99.00
F1-score (%)	97.70	97.70	98.00	98.00	100.00	100.00
Specificity (%)	99.21	96.81	98.00	98.00	100.00	99.00



(a) (b) (c)

Figure 13. The confusion matrices: (a) AtRD, (b) ViRD and (c) ReVi-RD.



(a) (b)

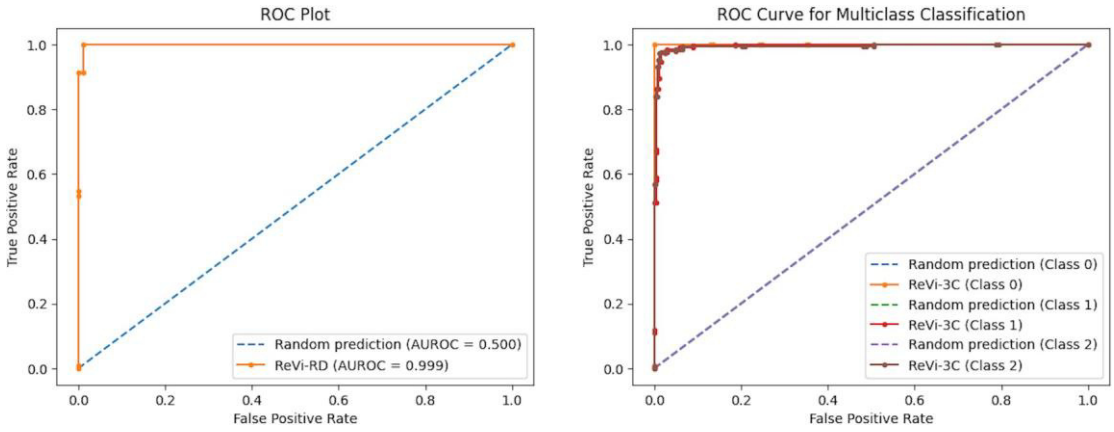
Figure 14. ROC curve for (a) ViRD and (b) ViR3C.

4.4 Diabetic Retinopathy Classification Performance

In the following experiment, we test the generalization capacity of the suggested models for the difficult task of classifying data into three different stages of severity in order to evaluate its potential.

Table 6 summarizes the evaluation metrics for staging RD into 3 classes. AtR3C and ViR3C offers a well-balanced performance across precision, recall and F1-score, as well as about 94% and 93% across

all metrics, respectively. ReVi-3C produced remarkable results, achieving an average of nearly 98% across all metrics and classes, including an area under the curve (AUC) of 99% per class, as shown in Figure 15b. This indicates that the model's predictions are balanced and reliable across the different performance measures.



(a) (b)

Figure 15. ROC curve for (a) ReVi-RD and (b) ReVi-3C.

Table 6. Performance evaluation of proposed models for 3-class DR classification (%).

Metric	AtR3C	ViR3C	ReVi-3C
Accuracy (%)	94.26	92.97	98.26
Precision (%)	94.41	93.77	98.43
Recall (%)	94.09	93.22	98.21
F1-score (%)	94.24	93.46	98.32
Specificity (Average) (%)	93.70	96.60	98.67

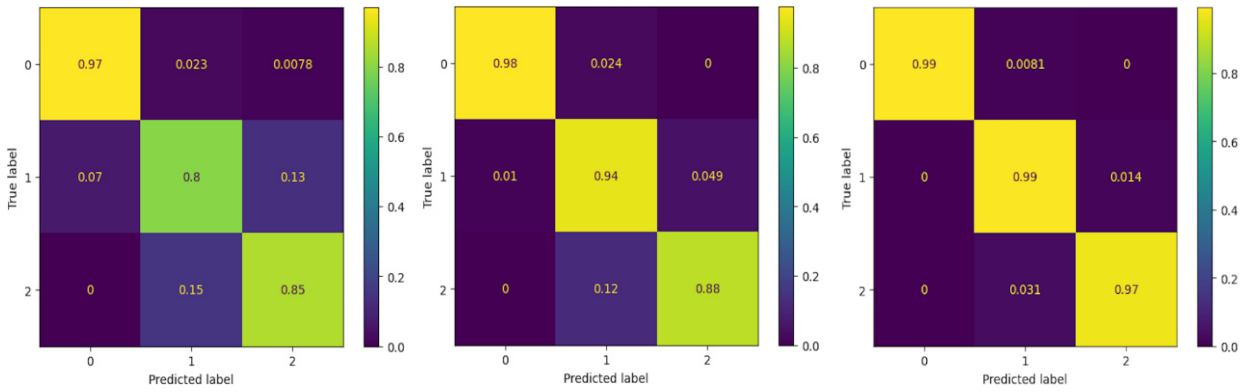
In order to evaluate the effectiveness of the suggested models (AtR3C, ViR3C and ReVi-3C), we examined the confusion matrices (see Figure 16), to provide details on the distribution of errors and classification accuracy across the three severity classes. As illustrated in Table 7, AtR3C model excels at identifying class 0 cases, achieving a precision of 97%, which means that nearly all predictions for this category are accurate. However, a specificity of 91.40% indicates that the model encounters difficulties with class 1. Specifically, 13% of cases are mislabeled as class 2 and 7% are incorrectly classified as class 0. Similarly, 15% of class 2 cases are mistakenly assigned to class 1. These patterns reveal a critical limitation: the model struggles to differentiate between adjacent severity levels, particularly distinguishing class 1 (moderate severity) from class 2 (high severity). This confusion suggests that AtR3C may lack the nuance needed to separate closely related categories, a gap that could impact its reliability in scenarios requiring precise severity staging. On the other hand, the ROC-curve in Figure 14b corresponding to class 0 lies very close to the top-left corner of the plot. This indicates that ViR3C is very accurate at detecting patients without DR.

The model demonstrated exceptional specificity of 99.5% for class 0 (healthy patients), minimizing false positives (0.5%) and thus avoiding misdiagnosis in unaffected individuals, which is essential for reliable screening. For class 1, specificity reached 92.8%, with 7.2% false positives, reflecting moderate difficulty in isolating this intermediate category. In contrast, class 2 (severe stage) has a high specificity of 97.5%, drastically limiting critical over-diagnosis and avoiding unwarranted invasive treatment.

For unhealthy cases, early-stage DR (class 1) is correctly identified in 94% of instances, though a 1% misclassification as healthy poses a risk of missed diagnoses, while advanced-stage DR (class 2) shows 88% accuracy, with 12% confused as early-stage DR, but none misclassified as healthy, highlighting robust performance for severe cases, but some overlap in staging severity. These results highlight the model's potential for accurately diagnosing early-stage DR and shows that the misclassification error mainly concerns stages 1 and 2.

Compared to AtR3C, ViR3C enhances the detection of healthy cases by reducing the misclassification rate of healthy individuals as non-healthy from 3% with AtR3C to 2% with ViR3C. This improvement highlights the power of ViTs in better detecting primitives across the entire set of images. We can decrease the errors by combining the strengths of the two architectures.

The hybrid ReVi-3C model dramatically outperforms its predecessors, AtR3C and ViR3C, achieving near-flawless classification across all severity levels: 99% precision for class 0 and class 1 and 97% for class 2, marking a substantial leap in accuracy. Misclassification errors are reduced to negligible levels, with only 3% of class 2 cases mistakenly labeled as class 1, while confusion between class 0 and class 1 is virtually eliminated. These results highlight the critical role of hybrid architectures in addressing multi-class challenges, where subtle inter-class differences demand precise discrimination.



(a) (b) (c)

Figure 16. The confusion matrix: (a) AtR3C, (b) ViR3C and (c) ReVi-3C.

Table 7. Class-wise performance of proposed models for 3-class DR classification (%).

	AtR3C			ViR3C			REVi-3C		
Metrics	**Class 0**	**Class 1**	**Class 2**	**Class 0**	**Class 1**	**Class 2**	**Class 0**	**Class 1**	**Class 2**
Precision (%)	98.00	82.00	86.00	98.00	91.00	93.00	100.00	98.00	98.00
Recall (%)	97.00	80.00	85.00	98.00	94.00	88.00	99.00	99.00	97.00
F1-score (%)	95.00	81.00	85.00	98.00	92.00	90.00	100.00	98.00	97.00
Specificity (%)	96.50	91.40	93.20	99.5	92.80	97.50	100.00	97.00	99.00

4.5 Results' Conclusion

The results obtained and their subsequent interpretation demonstrate that the proposed hybrid architectures (Revi-RD and Revi-3c) achieved remarkably high performance in both sensitivity and specificity. This success can be attributed to the effective exploitation of the complementary strengths of local feature extraction (by Resnet50) and global modeling of spatial dependencies (by ViTs).

5. Comparison of Our Approaches with the State-of-the-Art

To benchmark our approach, we compared our results with those of other state-of-the-art methods that have utilized transfer learning on the APTOS dataset for DR severity-level classification. Our models were benchmarked against Convolutional Neural Networks (CNNs) [32], [54], ensemble transfer learning [55], Supervised Contrastive Learning [34], a Deep Dual Branch model [56], Swin Transformer [50] and hybrid models combining Multiple Instance Vision Transformer (Milv4) [49] and Vision Transformer with Inception [47]. The comparison is carried out utilizing performance parameters including accuracy, precision, recall or sensitivity and F1-score across both binary and three-class classification tasks. All the methods illustrated in Table 8 are explained in the Related Works section. We can clearly say that our results are better and more enhanced than state-of-the-art results.

- **2-stage Classification**

AtRD model delivers a balanced performance (99.22% accuracy, 99.60% precision, 99.41% F1-score)

surpassing recent models, such as those of Shakibania et al. [56]. (98.50% accuracy) and Islam et al. [34] (98.36%). Athira et al. [55] achieved a slightly higher accuracy of 99.80%, as they also used an ensemble deep-learning approach with auto-tuning, but did not provide an F1-scor. In comparison, AtRD (99.22%) and ReVi-RD (99.55%) surpass nearly all previous works. However, the hybrid ReVi-RD model, with 99.55% accuracy, 99.51% precision and 99.54% F1-score, outperforms all existing approaches.

- **3-stage Classification**

AtR3C model did well in the 3-class-classification test, achieving accuracy, recall and F1-score values of 94.41%, 94.09% and 94.24%, respectively. Our results are somewhat superior to those of Athira et al. [55], who reported a slightly lower F1-score of 93.00%, but attained precision and recall of 94.00% each, noting that Athira did not report class performance. On the other hand, ViR3C attains an F1-score of 93.46%, demonstrating the potential of Vision Transformers (ViTs) in DR classification, though these models require more data than CNNs based on transfer learning. ReVi-3C, a hybrid architecture, achieves an impressive F1-score of 98.32%, representing an absolute improvement of 10.3% over Rao et al. and a 5.1% gain over Athira et al. This significant performance boost validates the effectiveness of hybrid models, where CNNs excel in localized feature extraction, while ViTs capture global contextual patterns. The importance of our method is underscored by the lack of research on the three-class classification of diabetic retinopathy (DR). Revi-3C's encouraging performance highlights its potential for DR detection, especially in its early stages, leading to better diagnostic results.

Table 8. Comparison of the proposed approaches with relevant previous works: binary and 3-stage classifications (unit %).

Architecture	Accuracy	Precision	Recall	F1-Score
Binary classification				
Esfahani [54]	86.00	85.00	86.00	85.00
Rao et al. [32]	96.56	97.00	97.00	96.56
Islam et al. [34]	98.36	98.37	98.36	98.37
Athira et al. [55]	99.80	99.00	99.00	99.00
Shakibania et al. [56]	98.50	97.61	99.46	/
Our AtRD	99.22	99.60	99.23	99.41
Dihin et al. [50]	96.00	/	98.00	96.00
Yang et al. [48]	93.2	/	86.9	/
Lian and Liul [47]	95.3	/	94.2	/
Our ViRD	97.73	97.72	97.73	97.73
Our ReVi-RD	**99.55**	99.51	**99.58**	**99.54**
3-class Classification				
Rao et al. [32]	/	88.00	88.00	88.02
Athira et al. [55]	94.00	94.00	93.00	
Our AtR3C	94.26	94.41	94.09	94.24
Our ViR3C	92.97	93.77	93.22	93.46
Our ReVi-3C	98.26	98.431	98.21	98.32

6. CONCLUSION

This study highlights the potential of Vision Transformers (ViTs) and hybrid architectures in advancing diabetic retinopathy (DR) classification, particularly for early detection. By simplifying the traditional five-stage DR classification into three classes—no DR, early DR (mild/moderate) and advanced DR (severe/proliferative), we reduced ambiguity between adjacent stages. To this end, we proposed three architectures: (1) a Resnet50-based model with Bayesian hyperparameter optimization (AtRD, AtR3C), (2) a fine-tuned Vision Transformer model (ViRD, ViR3C) and (3) a hybrid architecture (ReVi-RD, ReVi-3C) that combines the strengths of both approaches. Experimental

results show that while our architecture-based ViTs improve class differentiation, our hybrid model achieves superior accuracy and precision, demonstrating the advantage of integrating both local feature extraction and global attention mechanisms. This impressive result points to a high potential for accurate DR detection, which might greatly improve early diagnosis and care. However, several limitations should be noted. The use of the APTOS dataset alone for model training and evaluation may not fully represent the variety of fundus images encountered in real clinical settings. Consequently, it remains to generalize the models by training and evaluating on diverse datasets. Furthermore, the work does not fully address the difficulties of interpreting the models. It is essential to develop methods that enable clinicians to understand and trust the decisions made by the model. For future work, we aim to extend our model to five-stage DR classification to align with standard clinical grading. Additionally, we plan to enhance generalization by training and evaluating on diverse datasets, ensuring robustness across different populations and imaging conditions. Furthermore, we will investigate how to apply explainable AI approaches to improve the clarity of our model and encourage its application in medical environments.

ACKNOWLEDGEMENTS

This work was sponsored by the General Direction of Scientific Research and Technological Development, Ministry of Higher Education and Scientific Research (DGRSDT), Algeria.

REFERENCES

[1] E. Mehmet et al., "Diabetes Mellitus: A Review on Pathophysiology, Current Status of Oral Medications and Future Perspectives," Acta Pharmaceutica Sciencia, vol. 55, no. 1, pp. 61–82, 2017.

[2] J. Gu et al., "Recent Advances in Convolutional Neural Networks," Pattern Recognition, vol. 77, pp. 354–377, 2018.

[3] T. H. Fung et al., "Diabetic Retinopathy for the Non-ophthalmologist," Clinical Medicine, vol. 22, no. 2, pp. 112–116, 2022.

[4] D. J. Magliano et al., IDF Diabetes Atlas, 10th Edition, ISBN-13: 978-2-930229-98-0, 2022.

[5] F. Shaheen, B. Verma and M. Asafuddoula, "Impact of Automatic Feature Extraction in Deep Learning Architecture," Proc. of the 2016 IEEE Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8, Gold Coast, Australia, 2016.

[6] R. Adriman, K. Muchtar and N. Maulina, "Performance Evaluation of Binary Classification of Diabetic Retinopathy through Deep Learning Techniques Using Texture Feature," Procedia Computer Science, vol. 179, pp. 88–94, 2021.

[7] R. Rajkumar et al., "Transfer Learning Approach for Diabetic Retinopathy Detection Using Residual Network," Proc. of the 2021 6th IEEE Int. Conf. on Inventive Computation Technologies (ICICT), pp. 1189–1193, Coimbatore, India, 2021.

[8] S. Karthika et al., "Enhancing Diabetic Retinopathy Diagnosis with ResNet-50-based Transfer Learning: A Promising Approach," Annals of Data Science, vol. 11, no. 1, pp. 1–24, 2024.

[9] L. Dai et al., "A Deep Learning System for Detecting Diabetic Retinopathy across the Disease Spectrum," Nature Communications, vol. 12, no. 1, p. 3242, 2021.

[10] B. Tymchenko, P. Marchenko and D. Spodarets, "Deep Learning Approach to Diabetic Retinopathy Detection," arXiv preprint, arXiv: 2003.02261, 2020.

[11] P. Vashist et al., "Role of Early Screening for Diabetic Retinopathy in Patients with Diabetes Mellitus: An Overview," Indian Journal of Community Medicine, vol. 36, no. 4, pp. 247–252, 2011.

[12] K. Aggarwal et al., "Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning and Deep Learning," Iraqi J. for Comp. Sci. and Math., vol. 3, no. 1, pp. 115– 123, 2022.

[13] M. Bader Alazzam, F. Alassery and A. Almulihi, "Identification of Diabetic Retinopathy through Machine Learning," Mobile Information Systems, vol. 2021, no. 1, pp. 1–8, 2021.

[14] C. Mohanty et al., "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," Sensors, vol. 23, no. 12, p. 5726, 2023.

[15] C. Sharma and S. Parikh, "Comparison of CNN and Pre-trained Models: A Study," [Online], Available: https://www.researchgate.net/publication/359850786_Comparison_of_CNN_and_Pre-trained_models_A_Study, 2022.

[16] S. R. Salian and S. D. Sawarkar, "Melanoma Skin Lesion Classification Using Improved Efficientnetb3," Jordanian J. of Computers and Inform. Technol. (JJCIT), vol. 8, no. 1, pp. 45-56, 2022.

[17] I. Khoulqi and N. Idrissi, "Cervical Cancer Detection and Classification Using MRIS," Jordanian J. of Computers and Inform. Technol. (JJCIT), vol. 8, no. 2, pp. 141 – 158, 2022.

[18] I. Kandel and M. Castelli, "Transfer Learning with Convolutional Neural Networks for Diabetic Retinopathy Image Classification. A Review," Applied Sciences, vol. 10, no. 6, 2021.

[19] G. Selvachandran et al., "Developments in the Detection of Diabetic Retinopathy: A State-of-the-Art Review of Computer-aided Diagnosis and Machine Learning Methods," Artificial Intelligence Review, vol. 56, no. 2, pp. 915–964, 2023.
[20] S. D. Karthik, Maggie, "Aptos 2019 Blindness Detection," 2019.
[21] R. Casanova et al., "Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses," PLOS One, vol. 9, no. 6, p. e98587, 2014.
[22] T. M. Usman et al., "A Systematic Literature Review of Machine Learning-based Risk Prediction Models for Diabetic Retinopathy Progression," Artificial Intell. in Medicine, vol. 143, p. 102617, 2023.
[23] W. L. Alyoubi et al., "Diabetic Retinopathy Detection through Deep Learning Techniques: A Review," Informatics in Medicine Unlocked, vol. 20, p. 100377, 2020.
[24] S. Sengupta et al., "Ophthalmic Diagnosis Using Deep Learning with Fundus Images: A Critical Review," Artificial Intelligence in Medicine, vol. 102, p. 101758, 2020.
[25] H. Jiang et al., "Eye Tracking-based Deep Learning Analysis for the Early Detection of Diabetic Retinopathy: A Pilot Study," Biomedical Signal Processing and Control, vol. 84, p. 104830, 2023.
[26] R. Vij and S. Arora, "A Novel Deep Transfer Learning Based Computerized Diagnostic Systems for Multi-class Imbalanced Diabetic Retinopathy Severity Classification," Multimedia Tools and Applications, vol. 82, no. 22, pp. 34847–34884, 2023.
[27] P. Bijam and S. Deshmukh, "A Review on Detection of Diabetic Retinopathy Using Deep Learning and Transfer Learning-based Strategies," Int. Journal of Computer (IJC), vol. 45, no. 1, pp. 164–175, 2023.
[28] S. Z. Beevi, "Multi-level Severity Classification for Diabetic Retinopathy Based on Hybrid Optimization Enabled Deep Learning," Biomed. Signal Process. and Control, vol. 84, p. 104736, 2023.
[29] Z. Gu et al., "Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention," Comput. Intell. and Neurosci., vol. 2023, no. 1, p.1305583, 2023.
[30] H. E. Kim et al., "Transfer Learning for Medical Image Classification: A Literature Review," BMC Medical Imaging, vol. 22, no. 1, p. 69, 2022.
[31] O. Dekhil et al., "Deep Learning-based Method for Computer Aided Diagnosis of Diabetic Retinopathy," Proc. of the 2019 IEEE Int. Conf. on Imaging Systems and Techniques (IST), pp. 1–4, Abu Dhabi, UAE, 2019.
[32] M. Rao, M. Zhu and T. Wang, "Conversion and Implementation of State-of-the-Art Deep Learning Algorithms for the Classification of Diabetic Retinopathy," arXiv preprint, arXiv: 2010.11692, 2020.
[33] A. K. Gangwar and V. Ravi, "Diabetic Retinopathy Detection Using Transfer Learning and Deep Learning," Proc. of Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), vol. 1, pp. 679–689, 2021.
[34] M. R. Islam et al., "Applying Supervised Contrastive Learning for the Detection of Diabetic Retinopathy and Its Severity Levels from Fundus Images," Computers in Biology and Medicine, vol. 146, p. 105602, 2022.
[35] M. Oulhadj et al., "Diabetic Retinopathy Prediction Based on Deep Learning and Deformable Registration," Multimedia Tools and Applications, vol. 81, no. 20, pp. 28709–28727, 2022.
[36] S. S. Mondal et al., "EDLDR: An Ensemble Deep Learning Technique for Detection and Classification of Diabetic Retinopathy," Diagnostics, vol. 13, no. 1, p. 124, 2022.
[37] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint, arXiv: 2010.11929, 2020.
[38] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, no. 1, pp. 261–272, 2017.
[39] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint, arXiv: 2102.04306, 2021.
[40] X. Wang et al., "Transpath: Transformer-based Self-supervised Learning for Histopathological Image Classification," Proc. of 24th Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021), pp. 186–195, Part VIII 24, Strasbourg, France, 2021.
[41] J. Wu, R. Hu, Z. Xiao, J. Chen and J. Liu, "Vision Transformer-based Recognition of Diabetic Retinopathy Grade," Medical Physics, vol. 48, no. 12, pp. 7850–7863, 2021.
[42] N. J. Mohan, R. Murugan, T. Goel and P. Roy, "Vit-DR: Vision Transformers in Diabetic Retinopathy Grading Using Fundus Images," Proc. of the 2022 IEEE 10th Region 10 Humanitarian Technology Conf. (R10-HTC), pp. 167–172, Hyderabad, India, 2022.
[43] W. Nazih et al., "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-based Retina Images," IEEE Access, vol. 11, pp. 117546–117561, 2023.
[44] I. U. Khan et al., "A Computer-aided Diagnostic System to Identify Diabetic Retinopathy Utilizing a Modified Compact Convolutional Transformer and Low-resolution Images to Reduce Computation Time," Biomedicines, vol. 11, no. 6, p. 1566, 2023.
[45] T. Karkera et al., "Detecting Severity of Diabetic Retinopathy from Fundus Images: A Transformer Network-based Review," Neurocomputing, vol. 597, p. 127991, 2024.
[46] M. Oulhadj et al., "Diabetic Retinopathy Prediction Based on Vision Transformer and Modified

Capsule Network," Computers in Biology and Medicine, vol. 175, p. 108523, 2024.
[47] J. Lian and T. Liu, "Lesion Identification in Fundus Images *via* Convolutional Neural Network-vision Transformer," Biomedical Signal Processing and Control, vol. 88, p. 105607, 2024.
[48] Y. Yang, Z. Cai, S. Qiu and P. Xu, "A Novel Transformer Model with Multiple Instance Learning for Diabetic Retinopathy Classification," IEEE Access, vol. 12, pp. 6768 - 6776 2024.
[49] S. Yu et al., "Mil-vt: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification," Proc. of the 24th Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021), pp. 45–54, Part VIII 24, Strasbourg, France, 2021.
[50] R. A. Dihin et al., "Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet," Journal of Kufa for Mathematics and Computer, vol. 10, no. 2, pp. 167–172, 2023.
[51] S. V. M. Sagheer and S. N. George, "A Review on Medical Image Denoising Algorithms," Biomedical Signal Processing and Control, vol. 61, p. 102036, 2020.
[52] S. H. Abbood et al., "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model," IEEE Access, vol. 10, pp. 73079–73086, 2022.
[53] K. He et al., "Deep Residual Learning for Image Recognition," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, DOI 10.1109/CVPR.2016.90, 2016.
[54] M. T. Esfahani et al., "Classification of Diabetic and Normal Fundus Images Using a New Deep Learning Method," Leonardo Electronic J. of Practices and Techn., vol. 17, no. 32, pp. 233–248, 2018.
[55] T. Athira and J. J. Nair, "Diabetic Retinopathy Grading from Color Fundus Images: An Autotuned Deep Learning Approach," Procedia Computer Science, vol. 218, pp. 1055–1066, 2023.
[56] H. Shakibania et al., "Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy," Biomedical Signal Processing and Control, vol. 93, p. 106168, 2024.

ملخص البحث:

تعمل هذه الورقة على تقييم محوّلات الرّؤية وبِنى التّعلُّم العميق الهجينة من أجل تصنيف اعتلال الشّبكية لدى المصابين بمرض السُّكَّري، لمعالجة الغموض الّذي يكتنف التّمييز بين المراحل في الأنظمة التّقليدية. ففي حين تتفوّق الشّبكات العصبية الالتفافية في استخلاص السِّمات المحلّية في صوَر الشّبكية، فإنّ محوّلات الرّؤية توفر نمذجةً عالمية مثالية. وللاستفادة من نقاط القوّة تلك، نقترح بِنيةً هجينةً تجمع بين الشّبكات العصبية الالتفافية ومحوّلات الرّؤية بهدف التّصنيف الدّقيق لاعتلال الشّبكية لدى مرضى السُّكَّري. وقد قمنا ببناء ثلاثة نماذج لهذه الغاية وتقييمها: الأول يرتكز على الشّبكات العصبية الالتفافية وحدها، والثّاني يستند إلى محوّلات الرّؤية وحدها، والثالث نموذج هجين يجمع بينهما.

ولتقليل الغموض في التّمييز بين المراحل، قمنا بتقليل عدد مراحل اعتلال الشّبكية من خمس مراحل إلى ثلاث: لا اعتلال، واعتلال مبكّر (خفيف ومتوسط)، واعتلال متقدّم (شديد وقابل للانتشار). وقد تم تدريب النّماذج وتقييمها باستخدام مجموعة البيانات أبتوس (APTOS). ولدى مقارنة نتائج تقييم النّماذج الثّلاثة، تبين أنّ النّموذج الهجين يتفوق على النّموذجين الآخرين، محققاً دقّة تصنيف بلغت في معدّلها 98% في جميع أصناف الاعتلال الثلاثة. وقد حقّق النّموذج الهجين قيمةً ممتازة بلغت 99.5% في جميع مؤشّرات الأداء المتعلّقة بالتّصنيف الثّنائي (عدم وجود اعتلال؛ وجود اعتلال)، بينما بلغت تلك القيمة 98.3% في مؤشّرات الأداء المرتبطة بالتّصنيف الثّلاثي (لا اعتلال؛ اعتلال مبكّر؛ اعتلال متقدم). وخلاصة القول هي أنّ النّظام الهجين المقترح لتصنيف وجود أو عدم وجود اعتلال في الشّبكية لدى مرضى السُّكّري وتحديد درجة ذلك الاعتلال –إنْ وجد- تفوّق من حيث الأداء على الأنظمة التّقليدية الّتي تستخدم الشّبكات العصبية الالتفافية وتلك الّتي تستخدم طرقاً أخرى، حيث يعمل النّظام الهجين المقترح على التّقليل من اللّبْس في التّصنيف بين مراحل اعتلال الشّبكية، وذلك يبين أنّ لديه القدرة على التّصنيف الدّقيق للمراحل المختلفة لذلك الاعتلال.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).