DRCCT: Enhancing Diabetic Retinopathy Classification with A Compact Convolutional Transformer

Mohamed Touati [*] , Rabeb Touati , Laurent Nana , Faouzi Benzarti , Sadok Ben Yahia

Posted Date: 4 September 2024

doi: 10.20944/preprints202409.0303.v1

Keywords: AI; diabetic retinopathy; deep learning; DRCCT; classification; transformer

Article

DRCCT: Enhancing Diabetic Retinopathy Classification with A Compact Convolutional Transformer

Mohamed Touati [1,2,3], Rabeb Touati [3], Laurent Nana [1], Faouzi Benzarti [2], Sadok Ben Yahia [4]

[1] University of Western Brittany - UBO, Lab-STICC, Brest France; laurent.nana@univ-brest.fr
[2] University of Tunis, National High School of Engineering of Tunis; faouzi.benzarti@ensit.u-tunis.tn
[3] Faculty of Medicine of Tunis, Laboratory of Human Genetics; rabeb.touati@enit.utm.tn
[4] The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, say@mmmi.sdu.dk
* Correspondence: mohamed.touati@univ-brest.fr

Abstract: Diabetic retinopathy, a common complication of diabetes, is further exacerbated by factors such as hypertension and obesity. This study introduces the Diabetic Retinopathy Convolutional Transformer (DRCT) model, which combines convolutional and transformer techniques to enhance the classification of retinal images. The DRCT model achieved an impressive average F1 score of 0.97, reflecting its high accuracy in detecting true positives while minimizing false positives. Throughout 100 training epochs, the model exhibited strong generalization capabilities, achieving superior validation accuracy with minimal overfitting. On a newly evaluated dataset, the model attained precision and recall scores of 96.93% and 98.89%, respectively, indicating a well-balanced handling of false positives and false negatives. The model's ability to classify retinal images into five distinct diabetic retinopathy categories demonstrates its potential to significantly improve automated diagnosis and aid in clinical decision-making.

Keywords: AI; diabetic retinopathy; deep learning; DRCCT; classification; transformer

1. Introduction

Diabetes is a widespread metabolic condition that leads to multiple vascular complications throughout the body. The likelihood of eye-related problems escalates when diabetes is present alongside other health conditions like hypertension, obesity, and elevated cholesterol levels. learning (ML) techniques to enhance the detection and classification of DR. This condition harms the tiny blood vessels in the retina, resulting in a condition called diabetic retinopathy (DR). This frequent complication progressively damages these blood vessels, disrupting the retina's normal function. The damage can cause fluid to leak and blood vessels to become blocked, leading to significant vision loss or even blindness if left untreated. Diabetic retinopathy is the leading cause of global blindness, making early detection crucial. Emerging technologies, particularly artificial intelligence (AI), offer promising alternatives for cost-effective and efficient DR screening. Recent research [1]has focused on leveraging machine A study that reviewed various ML methods for DR detection highlighted the importance of early intervention and the potential of AI in providing scalable screening solutions. Figure 1 depicts a retinal fundus exam highlighting key features of diabetic retinopathy (DR). Visible are microaneurysms, small bulges within the retinal vessels, along with hemorrhages and exudates—indicators of bleeding and protein deposits. These findings are characteristic of non-proliferative diabetic retinopathy (NPDR), an early stage of the condition. DR is progressive, making early detection and treatment crucial to avoid significant vision loss. If left untreated, the disease may advance to proliferative diabetic retinopathy (PDR), marked by abnormal blood vessel growth.

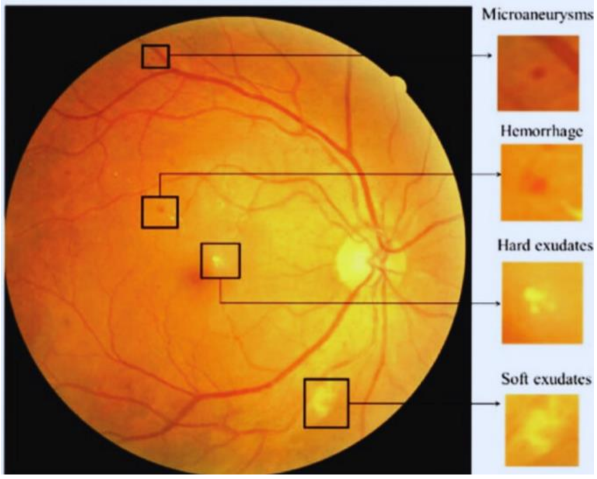


Figure 1. Diabetic Retinopathy: Key Features.

The study conducted a bibliometric analysis using data from Scopus and Web of Science to explore different ML styles used in DR diagnosis, combining quantitative and qualitative analyses to offer insights into image segmentation methods, datasets, and ML approaches, including traditional and deep learning techniques.Advances in artificial intelligence (AI) present new ways to enhance disease detection and management. A 178 reviewed studies [2] on DR screening systems using AI techniques, highlighting the urgent need for automated, reliable solutions due to the global rise in DR patients. The review spans publications from January 2014 to June 2022, discussing various AI, machine learning (ML), and deep learning (DL) tools used for DR detection. A key focus is on the comparison between custom-built convolutional neural networks (CNNs) and those employing transfer learning with established architectures like VGG, ResNet, or AlexNet. While creating a CNN from scratch requires significant time and resources, transfer learning offers a quicker alternative. However, studies indicate that custom CNN architectures often outperform those using existing structures. This distinction warrants further research. The survey also explores feature extraction techniques, which enhance model performance by reducing feature vector size and computational effort. Publicly available datasets are analyzed, along with performance metrics crucial for evaluating the accuracy and effectiveness of DR detection systems. The review identifies a gap in technologies capable of predicting all DR stages and detecting various lesions, highlighting the need for advanced solutions to improve patient outcomes and prevent vision loss. Future research should consider emerging concepts like transfer learning, ensemble learning, explainable AI, multi-task learning, and domain adaptation to enhance early DR detection.

Recent developments in Deep Learning, especially with Vision Transformers (ViTs), have demonstrated significant potential in the field of medical imaging. The number of publications on ViTs surged to 19 by the end of 2022, underscoring their ability to enhance medical image analysis [3]. ViTs improve both the accuracy and speed of analyzing retinal images, which is crucial for early diagnosis and timely intervention. Our project leverages these advancements by incorporating ViTs into our AI tools for detecting and managing diabetic retinopathy. This strategy aims to equip healthcare professionals with advanced tools for more effective diagnosis and treatment of diabetic retinopathy, ultimately aiding in the preservation of patients' vision.

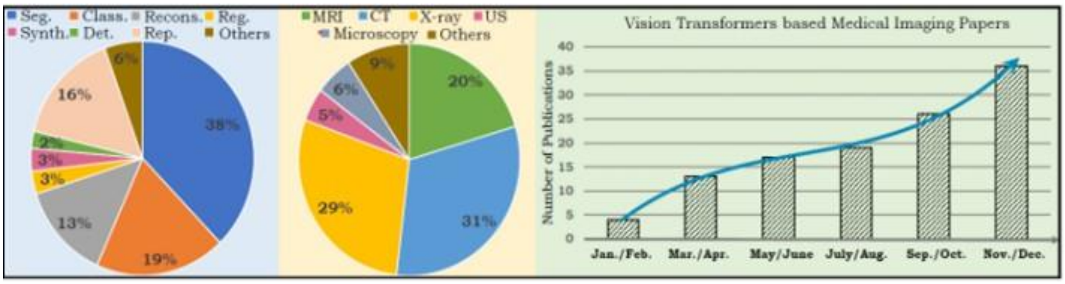


Figure 2. Trends in Medical Imaging Research: Modalities and Fields [3].

Our project introduces a novel AI model based on Tronsformer for detecting diabetic retinopathy, focusing on overcoming challenges such as human subjectivity and limited access to traditional screening methods. With the rise in diabetes globally, early detection and management of diabetic retinopathy are crucial to prevent severe complications and vision loss. To address this need, we are developing a platform in collaboration with Pixemantic and doctors that integrates advanced AI models to enhance the detection of different stages of the disease. This project simulates retinal images under controlled conditions to evaluate the effectiveness of these AI models.

2. Literature Review

In recent years, the demand for precise diagnosis of Diabetic Retinopathy (DR) has received considerable attention, prompting the development of numerous Computer-Aided Diagnosis (CAD) methods designed to aid clinicians in interpreting fundus images. Deep learning algorithms have particularly stood out due to their exceptional ability to automatically extract and classify features. For example, Sheikh and Qidwai [4] applied the MobileNetV2 architecture on a different dataset, utilizing transfer learning to achieve a remarkable 90.8% accuracy in diagnosing DR and 92.3% accuracy in identifying referable diabetic retinopathy (RDR) casesIn [5], the researchers tackled the problem as a binary classification task, attaining an impressive 91.1% accuracy on the Messidor Dataset and 90.5% on the EyPacs Database. These results underscore the method's strong potential for application in clinical environments. Moreover, the study in [6] proposed a multi-channel Generative Adversarial Network (GAN) with semi-supervised learning for assessing diabetic retinopathy (DR). The model tackles the issue of mismatched labeled data in diabetic retinopathy (DR) classification through three primary mechanisms: a multi-channel generative approach to produce sub-field images, a multi-channel Generative Adversarial Network (GAN) with semi-supervised learning to effectively utilize both labeled and unlabeled data, and a DR feature extractor designed to capture representative features from high-resolution fundus images. In their study [4], Touati et al. began the retinopathy workflow by converting images into a hierarchical data format, which included steps such as pre-processing, data augmentation, and training. The Otsu method was employed for image cropping, specifically to isolate the circular-colored retinal regions. Normalization was then applied, where the minimum pixel intensity was subtracted, and the result was divided by the average pixel intensity, bringing the pixel values into the 0 to 1 range. Contrast enhancement was accomplished using adaptive histogram equalization filtering, specifically with CLAHE. In [7], M. Touati et al. presented an approach that combines image processing with transfer learning techniques. The advanced image processing steps are designed to extract richer features, improving the quality of subsequent analysis. Transfer learning, using the Xception model, speeds up the training process by utilizing pre-existing knowledge.These combined techniques resulted in high training accuracy (92%) and test accuracy (88%), demonstrating the effectiveness of the proposed method. In a separate study, Yaakoob et al. [8] developed a method for detecting and grading diabetic retinopathy by merging ResNet-50 features with a Random Forest classifier. This approach leverages features from ResNet-50’s average pooling layer and highlights the role of specific layers in improving performance. ResNet helps overcome issues like vanishing gradients, enabling effective training of deeper networks. In article [9], researchers used feature extraction to identify anomalies in retinal images, allowing for quick diabetic retinopathy (DR) detection on a scale of 0 to 4. Various classification algorithms were tested, with the Naïve Bayes Classifier achieving 83% accuracy.In [10], Toledo-Cortés et al. presented DLGP-DR, an advanced deep learning model that improved classification and ranking of diabetic retinopathy (DR) using a Gaussian process. DLGP-DR outperformed previous models in accuracy and AUC scores, providing enhanced insights into misclassifications [11]. Experiments on the Messidor dataset demonstrated that the proposed model outperforms other notable models [11,12], in terms of accuracy, AUC, sensitivity, and overall performance, even with only 100 labeled samples. The approach utilizes deep learning with a CNN and attention network, achieving Kappa scores of 0.857 and 0.849, and sensitivity rates of 0.978 and

0. 960.In [13], TOUATI et al. introduced a ResNet50 model integrated with attention mechanisms, marking a significant advancement in diabetic retinopathy (DR) detection. The model achieved a training accuracy of 98.24% and an F1 Score of 95%, demonstrating superior performance compared to existing methods. The approach described in [14], named TaNet, leverages transfer learning for classification and has shown excellent results on datasets such as Messidor-2, EYEPACS-1, and APTOS 2019. The model achieved impressive metrics, including 98.75% precision, 98.89% F1-score, and 97.89% recall, outperforming current methods in terms of accuracy and prediction performance. In [15], four scenarios using the APTOS dataset were tested with HIST, CLAHE, and ESRGAN. The CLAHE and ESRGAN combination achieved the highest accuracy of 97.83% with a CNN, matching experienced ophthalmologists. This underscores the value of advanced preprocessing in improving DR detection and suggests further research on larger datasets could be beneficial. In a manner similar to [17], which introduced a novel ViT model for predicting diabetic retinopathy severity using the FGADR dataset, [16] underscores the potential of Vision Transformers in advancing diagnostic accuracy and performance in medical imaging tasks. The study in [18] presents DR-CCTNet, a modified transformer model designed to improve automated DR diagnosis. Tested on diverse fundus images from five datasets with varying resolutions and qualities, the model utilized advanced image processing and augmentation techniques on a large dataset of 154,882 images. The compact convolutional transformer was found to be the most effective, achieving 90.17% accuracy even with low-pixel images. Key contributions include a robust dataset, innovative augmentation methods, improved image quality through pre-processing, and model optimization for better performance with smaller images.In [19], a new deep learning model, Residual-Dense System (RDS-DR), was developed for early diabetic retinopathy (DR) diagnosis. This model combines residual and dense blocks to effectively extract and integrate features from retinal images. Trained on 5,000 images, RDS-DR achieved a high accuracy of 97% in classifying DR severity. It outperformed leading models like VGG16, VGG19, Xception, and InceptionV3 in both accuracy and computational efficiency. Beraber [20] presents a novel approach for detecting and classifying diabetic retinopathy using fundus images. The method employs a feature extraction technique known as "Uniform LocalBinary Pattern Encoded Zeroes" (ULBPEZ), which reduces feature size to 3.5% of its original size for more compact representation. Preprocessing includes histogram matching for brightness standardization, median filtering for noise reduction, adaptive histogram equalization for contrast enhancement, and unsharp masking for detail sharpening. Nafseh Ghafar et al. [22] emphasize that deep learning (DL) algorithms excel in medical image analysis, especially for fusion, segmentation, registration, and classification tasks. Among machine learning (ML) and deep learning (DL) techniques, support vector machines (SVM) and convolutional neural networks (CNN) are particularly noted for their effectiveness.Yasashvini R et al. [21] investigated the use of convolutional neural networks (CNN) and hybrid CNNs for diabetic retinopathy classification. They developed several models, including a standard CNN, a hybrid CNN with ResNet, and a hybrid CNN with DenseNet. The models achieved accuracy rates of 96.22%, 93.18%, and 75.61%, respectively. The study found that the hybrid CNN with DenseNet was the most effective for automated diabetic retinopathy classification. Nafseh Ghafar et al. [22] highlight that healthcare's vast data is ideal for Deep Learning (DL) and Machine Learning (ML) advancements. Medical images from various sources are key for improving analysis. To enhance image quality for CAD systems in diabetes detection, techniques like denoising, normalization, bias field correction, and data balancing are used. These methods reduce noise, standardize intensity, correct intensity variations, and address class imbalances, respectively, to improve image analysis. Yaoming Yang et al. [23] examined the advancement of Transformers in NLP and CV, highlighting the 2017 introduction of the Transformer, which improved NLP by capturing long-range text dependencies. Their machine learning process involves resizing retinal images to 448 x 448 pixels, normalizing them, and dividing them into 16 x 16-pixel patches with random masks. These patches are processed by a pre-trained Vision Transformer (ViT) to extract features, which are then decoded, reconstructed, and used by a classifier to detect diabetic retinopathy (DR). The study found that using Vision Transformers (ViT) with Masked Autoencoders (MAE) for pre-training on over 100,000 retinal images resulted in better DR detection than pre-

training with ImageNet, achieving 93.42% accuracy, 0.9853 AUC, 0.973 sensitivity, and 0.9539 specificity.More recently, in 2021, Nikhil Sathya et al. [24] introduced an innovative approach by combining Vision Transformers (ViT) with convolutional neural networks (CNNs) for medical image analysis. Jianfang Wu et al. [25] highlighted the importance of attention mechanisms in natural language processing, noting that transformers, which eschew traditional convolutional layers for multi-head attention, offer advanced capabilities. [28] Although CNN have proven effective in grading diabetic retinopathy by efficiently extracting pixel-level features, the emergence of transformers offers potential benefits in this field. Integrating CNNs with Vision Transformers (ViTs) has shown to be more effective than relying solely on pure ViTs, as CNNs are limited in handling distant pixel relationships, while ViTs perform exceptionally well in complex tasks like dense prediction and detecting tiny objects. However, ViTs are still considered a black box due to their opaque internal processes, highlighting the need for further research to create explainable ViT models or hybrid CNN-ViT models for diabetic retinopathy classification and similar applications.

3. Transformer

Transformers are increasingly used in natural language processing and medical imaging due to their ability to capture contextual information and long-term relationships. Transformers have been extensively integrated into various fields, including natural language processing and medical imaging. Their ability to capture contextual information and long-term relationships is particularly beneficial for applications such as image segmentation, classification, and disease detection, enhancing diagnostic accuracy and facilitating medical decision automation. According to Shamshad et al. [3], Figure 9 shows a notable increase in research publications on Vision Transformers (ViT) applied to medical imaging. Since January 2020, there has been a significant rise in publications, reaching 19 by the end of 2022. This trend reflects growing interest in ViTs and their revolutionary potential in medical image analysis. ViTs have diverse applications in medicine, including image segmentation, reconstruction, and classification.

3.1. ViT: Challenging CNNs and RNNs in Image Classification

The Vision Transformer (ViT) represents a significant breakthrough in artificial intelligence applied to image recognition, emerging as a promising alternative to convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in image classification tasks. Developed by researchers at Google Brain, the ViT takes an innovative approach by segmenting images into patches and processing them through a transformer-based encoding architecture. This allows the model to effectively capture global dependencies using self-attention mechanisms. Unlike CNNs, which focus on local patterns in a hierarchical manner, and RNNs, which handle sequential information, the ViT processes local features within patches while simultaneously considering the entire image, thus offering a global receptive field. This approach surpasses the local and sequential processing capabilities of CNNs and RNNs. Additionally, the parallelizable nature of the transformer's architecture enhances the scalability of ViT, giving it an edge over other models whose scalability is constrained by their sequential data processing methods.

As shown in Table 1, ViT architectures have outperformed CNNs in complex tasks such as dense prediction and tiny object detection by utilizing advanced internal representations of visual data. Despite these advancements, the internal representations of ViTs are often opaque, treating the model as a "black box." To improve the understanding and interpretation of ViT models, especially in medical image analysis and classification, developing new visualization layers is essential. This research aims to enhance the explainability of vision transformers for more effective applications in medical imaging [29].

Table 1. Comparison of Neural Network Architectures: CNNs, RNNs, and ViTs.

Aspect	CNNs	RNNs	ViTs

Architecture	Convolutional layers	Sequential recurrent layers	Transformer Encoder with self-attention
Data Processing	Local patterns, spatial hierarchies	Sequential information	Dependencies, global integration
Feature Learning	Local features, sequential learning	Global features, entire sequence	Local integration into patches, global integration
Receptive Field	Local	Local (sequential)	Global
Feature Engineering	More manual, learns from data	More manual, learns from data	Less manual, learns from data
Scalability	Average	Low (sequential processing)	High (parallel processing)

3.2. Main Components of a Vision Transformer

The Vision Transformer (ViT) is a specialized adaptation of the original Transformer architecture designed for image classification tasks. It starts by dividing an image into a grid of 2D patches, each with a specific resolution. These patches are then flattened and projected into a higher-dimensional space to create "patch embeddings." To capture the spatial relationships between patches, ViT includes learnable token embeddings, akin to the [CLS] tokens used in BERT, which represent the entire image context. Positional encodings are added to preserve the spatial arrangement of the patches [26]. ViT functions as a traditional transformer encoder, processing sequences of these embeddings through self-attention and feedforward layers. The final output from the encoder is then passed through a multi-layer perceptron (MLP) head for classification. This structure allows ViT to effectively analyze and classify images by considering the contextual relationships among the patches. This section explores the fundamental concepts of ViT, focusing on its attention mechanism and the various functional blocks depicted in the Figures 3

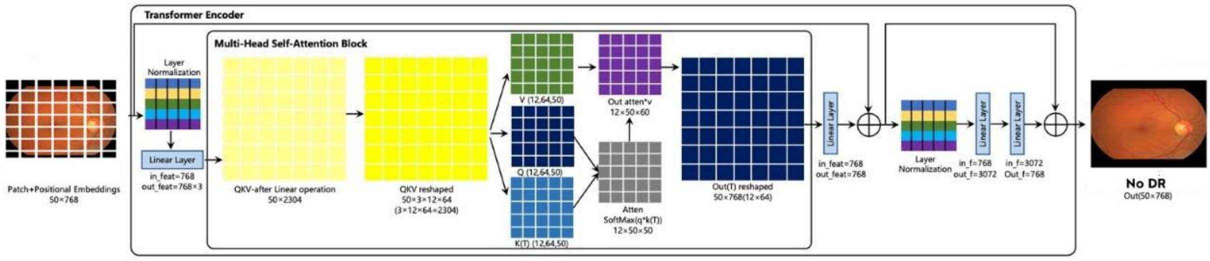


Figure 3. Transformer Encoder Block in Vision Transformer with Multi-Head Self-Attention Module.

3.2.1. Transformer Encoder

The Vision Transformer (ViT) encoder is composed of alternating layers of Multi-Head Attention (MHA) blocks and Multi-Layer Perceptron (MLP) blocks. Before each transformation block, layer normalization is applied, and residual connections are added after each block. These residual connections (also known as "skip connections") provide alternate pathways for data, allowing it to bypass certain layers and reach deeper parts of the model more directly. Layer normalization is a technique used to standardize the distribution of inputs to each layer of the model, improving learning speed and generalization accuracy. It involves centering and rescaling the input vector representation to ensure consistency in the input size for the normalization layer. Unlike traditional

Transformer blocks that have both encoding and decoding layers, the Vision Transformer only has an encoding layer. The output of the transformer encoder is then sent to the MLP head, which performs class classification based on the image representations learned from the class labels in the final layer [26].

3.2.2. Patch Embedding

To address memory constraints, images are divided into smaller patches for sequential processing. Each patch is converted into a feature vector, drawing on the embedding concept used in Vision Transformers (ViT)[27]. These vectors are visualized in an embedding space, where similar features group together, aiding in classification. Figure 3 (a part)demonstrates this process, with embedding layers being refined during training. This approach, particularly in retinal imaging, combines positional encoding with feature embedding to ensure accurate feature selection.

3.2.3. Position Encoding

In architectures that use patch embedding, a key challenge is the limited knowledge of each patch's position, making it difficult to establish relationships between them. Transformers address this issue with positional embedding, which preserves the positional information of tokens within a sequence. This is particularly important in fields like medical imaging, where precise feature identification is critical. Unlike traditional methods, transformers use positional embeddings, which are learned during training, to incorporate positional information. In vision transformers, these embeddings are essential because image patches do not naturally contain spatial information. Positional embeddings are combined with patch embeddings to encode the location of each patch in the image, linking feature vectors to their positions in the sequence. Positional encoding is usually implemented with sine and cosine functions at different frequencies for each embedding dimension. These values are then merged with feature vectors to create a new vector that represents both the feature and its position.

3.2.4. Attention Mechanism

Attention mechanisms, inspired by human visual focus, improve deep learning models by emphasizing the most relevant parts of an image. This selective emphasis helps the model capture crucial contextual information while ignoring noise, enhancing the accuracy and efficiency of tasks like image classification, object detection, and semantic segmentation. There are two main types of attention mechanisms: self-attention, which analyzes relationships within a sequence, and multi-head attention, which applies self-attention across multiple subspaces. The core function of attention mechanisms is to capture dependencies between elements in a sequence, regardless of their position.

3.2.4. Self-Attention

The self-attention mechanism is fundamental to the Transformer's architecture, enabling it to model long-term dependencies in a sequence. It generates a representation for each sequence element by considering the influence of all other elements. This is done by calculating similarity scores between pairs of elements, which are then converted into attention weights using a softmax function. These weights help create a weighted sum of the original element representations, capturing the sequence's global context. The self-attention mechanism involves three key components: the query (Q), the key (K), and the value (V). The query is the element being contextualized, the key is used to determine relevance, and the value is the element weighted by the attention score to produce the final output.

3.2.5. Multi-Head Self-Attention Mechanism

The multi-head attention mechanism in Transformers uses multiple parallel self-attention "heads," each focusing on different data aspects. These heads apply distinct transformations to the input, highlighting unique features. Their outputs are then combined and further processed to enhance the model's understanding of the data. The classification head in the Vision Transformer converts the encoder's output into class probabilities. It typically involves a multi-layer perceptron (MLP) or a linear layer, which processes and flattens the patch embeddings, applies dropout to avoid overfitting, and then predicts the image class.

3.3. Compact Convolutional Network

In our study, we introduce the Compact Convolutional Transformer (CCT) as a highly efficient model for classifying and detecting the stages of diabetic retinopathy. Unlike other transformer-based models, CCT excels in performance in the work of [29] on smaller datasets while also significantly reducing computational costs and memory usage. This efficiency challenges the conventional notion that transformers require vast computational resources, making them accessible even in resource-limited settings. The CCT's ability to operate effectively with limited data highlights its potential for broader application in various scientific domains where data availability is often constrained, thereby extending the reach and impact of machine learning research.

3.4.1. Convolutional Tokenization

Convolutional tokenization serves as the initial step in the CCT architecture defined in the Figure 4, where regions of interest within retinal images are segmented using convolutional layers. These layers are configured with specific parameters such as kernel size, stride, and padding, which dictate how the images are divided into patches.For an image with height 112, width 112, and number of channels 3, convolutional tokenization is employed to extract features from each patch. The process can be represented by the following sequential operation:

$$x_0 = \mathrm{AveragePool}(\mathrm{ReLU}(\mathrm{conv2D}(x))) \qquad (1)$$

In the DRCCT (Diabetic Retinopathy Compact Convolutional Transformer) model, four different filters—16, 32, 64, and 128—are used within the CCT tokenizer. These filters determine the number of output channels or feature maps produced by the convolutional layer. By adjusting the size and quantity of patches through the use of various filters, the model achieves a balance between the detail within patches and the overall sequence length generated.

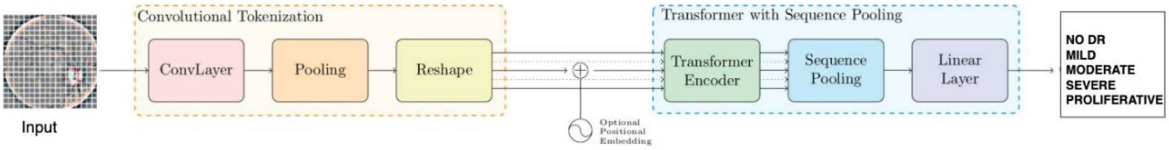


Figure 4. Compact Convolutional Network Architecture.

The Compact Convolutional Transformers (CCT) architecture combines a convolutional tokenizer, SeqPool, and a transformer encoder. CCT variants are denoted by the number of transformer encoder layers and convolutional layers, such as CCT-7/3x2, which signifies a model with 7 transformer encoder layers and a 2-layer convolutional tokenizer with a 3×3 kernel size[29].

3.4.2. Transformer Encoder

Following convolutional tokenization, the sequences are processed through a series of transformer blocks in the CCT architecture. Each transformer block includes two main components: a Multi-Head Attention (MHA) layer and a Multi-Layer Perceptron (MLP) block. The patches are encoded using layer normalization, MHA, and MLPs with ReLU activation and dropout. Key parameters such as the number of transformer layers, output channels, hidden units, and dropout rates are carefully defined to optimize the model’s performance. Stochastic depth is employed as a regularization method, which involves applying residual branches from transformer blocks before the residual connections during training. This technique reduces the network's effective depth, improving generalization and reducing the risk of overfitting. The output of the transformer encoder is a tensor containing encoded patch features, which is then prepared for further processing and classification.

3.4.3. Sequence Pooling

In traditional transformer models like ViT and BERT, global average pooling is used to condense the output token sequence into a single class index. The newer "sequence pooling" approach, however, employs an attention-based mechanism to retain essential information from various parts of the input image. This method enhances model performance without extra parameters and slightly reduces computational demand. The sequence pooling process begins by transforming the output sequence of the transformer encoder:

$$x^{L}= f\ (x^{0}) \in \mathrm{Rb,n,d} \quad (2)$$

where $x^{L}$is the output from layer LLL of the transformer encoder, bbb is the batch size, nnn is the sequence length, and d is the total embedding dimension. This output is then processed through a linear layer:

$$x' = \mathrm{softmax}(g\ (x^{L})^{T}) \in \mathrm{Rb,1,d} \quad (3)$$

where x′ contains the importance weights for the tokens. These weights are applied to the output sequence to produce the final weighted output:

$$z = x^{L} \odot x' \qquad (4)$$

The result, z, is a weighted and flattened output used for classification purposes.

3.4.4. Classification Tasks

In the final stage of the CCT model, dense layers are employed for the classification of diabetic retinopathy stages. The final dense layer typically outputs class probabilities for multi-class classification tasks or a single value for binary classification. Dense neural networks are particularly effective at learning complex patterns from input data, making them a popular choice in machine learning and deep learning applications, especially for tasks involving image classification.

4. Work Done

4.1. Data Undestading

The data used in this study comes from Kaggle, a well-known platform for data science research and competitions. Specifically, we utilized the diabetic retinopathy dataset, which contains five categories of images for fundus imaging. Our dataset consists of 6399 images, which we divided into 80% for training and 20% for testing. Each image is labeled with the severity level of the disease: No_DR, Mild, Moderate, Severe, and Proliferated. Although information such as the patient's age is not included, the samples are individual color (RGB) images of 176 x 208 pixels. Table 2 shows the distribution of the dataset, indicating the number of images per category.

The data used in this study comes from Kaggle, a well-known platform for data science research and competitions. Specifically, we utilized the diabetic retinopathy dataset, which contains five categories of images for fundus imaging. Our dataset consists of 6399 images, which we divided into 80% for training and 20% for testing. Each image is labeled with the severity level of the disease: No_DR, Mild, Moderate, Severe, and Proliferate_DR. Although information such as the patient's age is not included, the samples are individual color (RGB) images of 176 x 208 pixels. Table 2 shows the distribution of the dataset, indicating the number of images per category.

Table 2. Distribution of the DR dataset.

Classes	Train set	Test set
No DR	2192	549

Mild	592	148
Moderate	1518	380
Proliferative DR	472	118
Severe	284	72

4.2. Image Preprocessing

4.2.1. feature Extraction

In image preprocessing, we ensured data quality and consistency by extracting images while preserving visual details. We applied resizing and pixel normalization for uniform scaling and effective data management. To enhance image representation and feature extraction, we used convolution and resizing techniques to emphasize key patterns and structures in retinal images presented in the Figure 5. Convolutional layers in our model were crucial for automatically detecting important visual patterns, which aids in precise feature extraction and accurate diabetic retinopathy analysis. Additionally, these layers helped reduce dimensionality, providing a more informative and compact data representation.

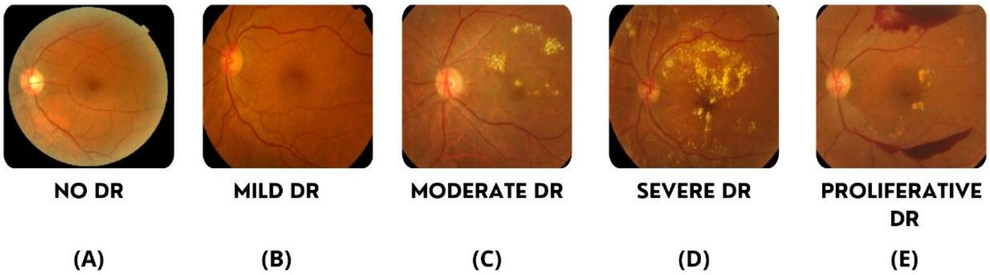


Figure 5. Diabetic Retinopathy Classes.

4.2.1. Noise Reduction

Image preprocessing plays a crucial role in enhancing the quality and effectiveness of diabetic retinopathy (DR) detection. The process begins with **grayscale conversion**, which simplifies the image data by reducing it from RGB to a single channel of intensity values. This step is essential for focusing on the critical features necessary for DR detection, achieved through the formula (5):

$$I_{\text{gray}} = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \qquad (5)$$

The result is a single-channel image that minimizes computational complexity while retaining key visual information. Following this, CLAHE (Contrast Limited Adaptive Histogram Equalization) is applied to the grayscale image. CLAHE enhances the local contrast of the image, making subtle DR features, such as microaneurysms, more visible. This technique improves local contrast without amplifying noise, thereby highlighting critical features more effectively.

To further refine the image, Gaussian smoothing is employed. This technique involves applying a Gaussian filter to the CLAHE-enhanced image to reduce noise while preserving important edges. The Gaussian filter is represented by the formula (6):

$$I_{median} = I_{CLAHE} * G_{\sigma} \qquad (6)$$

where Gσ denotes the Gaussian kernel with standard deviation σ. This results in a smoothed image that facilitates better feature detection and segmentation. Finally, median filtering is used to address any remaining noise, particularly salt-and-pepper noise, while maintaining edge integrity. By applying a median filter (7) to the smoothed image:

$$I_{median} = \text{MedianFilter}(I_{smoth}). \qquad (7)$$

The outcome is a further noise-reduced image that preserves fine details and edges, thereby enhancing the accuracy of subsequent feature extraction. Together, these preprocessing steps ensure that the image is optimally prepared for the detection of diabetic retinopathy, enhancing both feature visibility and overall model performance.

4.2.3. Data Augmentation

This project aimed to enhance a diabetic retinopathy detection model by addressing the challenges posed by limited, imbalanced, and noisy fundus image datasets. We used data augmentation methods such as rotation, resizing, flipping, cropping, shifting, and noise addition to increase the quantity, diversity, and quality of the data. By providing more varied and realistic data, we aimed to enhance the model's generalization, reduce overfitting, and boost its ability to accurately detect diabetic retinopathy.

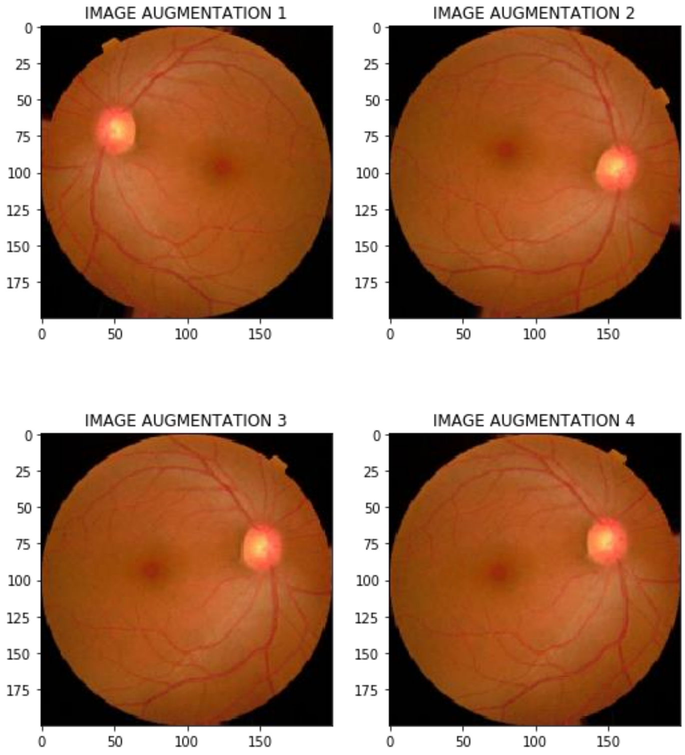
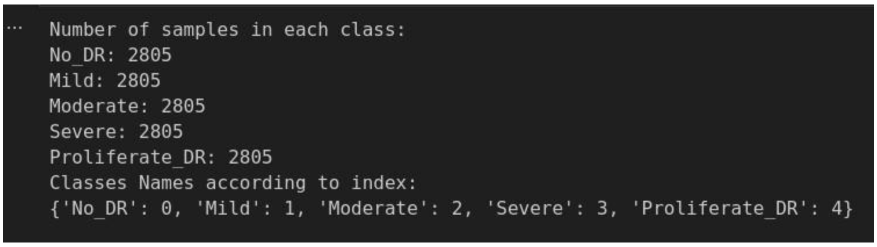


Figure 6. Data Augmentation Operation.

4.2.4. Data Balancing

To ensure fair and accurate diabetic retinopathy classification, it's crucial to address class imbalance. When certain classes dominate the dataset, models can become biased, performing well on majority classes but poorly on minority ones. By using techniques like the RandomOverSampler from the imbalanced-learn library, as mentioned in the figure below, we can balance the dataset by duplicating samples from minority classes. This process ensures each class has an equal number of samples : 2805 allowing the model to learn and predict across all categories with greater accuracy and fairness.



```
...  Number of samples in each class:
     No_DR: 2805
     Mild: 2805
     Moderate: 2805
     Severe: 2805
     Proliferate_DR: 2805
     Classes Names according to index:
     {'No_DR': 0, 'Mild': 1, 'Moderate': 2, 'Severe': 3, 'Proliferate_DR': 4}
```

Figure 7. Capture of data Balancing Result.

4.3. Modeling Bulding

The DRCCT model leverages a sophisticated multi-step pipeline to achieve high classification accuracy for diabetic retinopathy. It begins with the collection and preprocessing of retinal images, employing median filtering and CLAHE to enhance image quality. To address class imbalance, RandomOverSampler is used to balance the dataset to 2805 samples per class, complemented by extensive data augmentation using nine transformations.

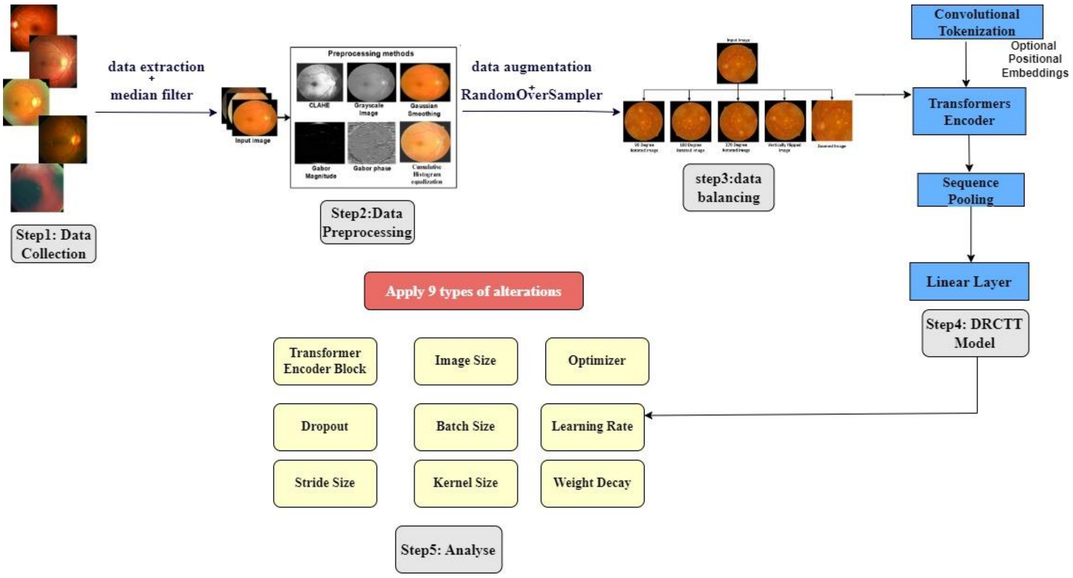


Figure 8. Our DRCCT Workflow.

The model incorporates convolutional tokenization with four different filter sizes (16, 32, 64, 128) to generate patch sequences. Positional embeddings are added to maintain spatial information. The architecture includes the following components:

Input Layer: (None, 112, 112, 3) — Accepts retinal images of size 112x112 with 3 color channels.

CCT Tokenizer: (None, 4, 120) — Performs convolutional tokenization, producing sequences of 120-dimensional patches from the image.

tf.operators_add: (None, 4, 120) — Applies element-wise addition to integrate information from different sources.

Layer Normalization: (None, 4, 120) — Normalizes the tokenized sequences to stabilize training and improve convergence.

Multi-head Attention (AMH): (None, 4, 120) — Utilizes multi-head attention to capture complex dependencies and relationships between different parts of the image.

Stochastic Depth: (None, 4, 120) — Employs stochastic depth regularization to enhance model generalization by randomly dropping layers during training.

Add: (None, 4, 120) — Combines information from previous layers or sources to refine feature representation.

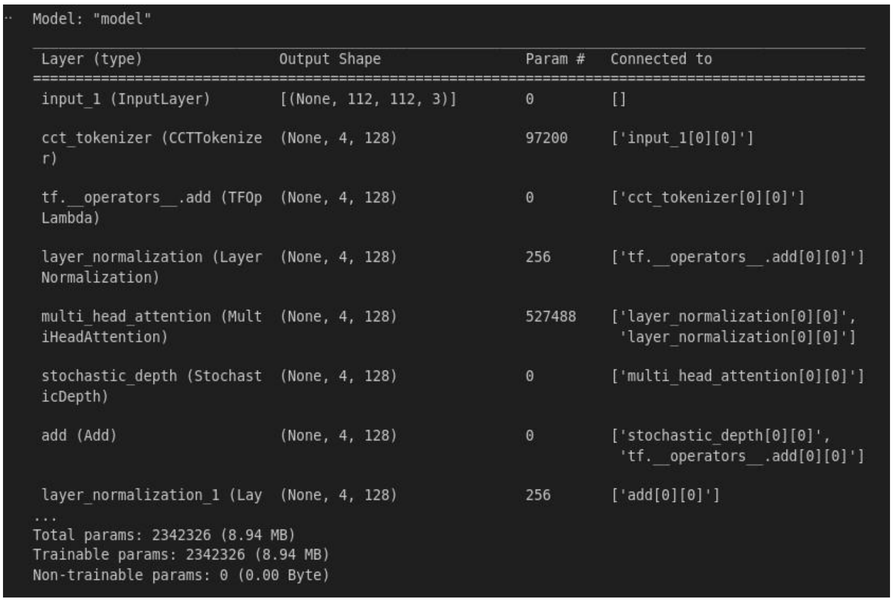
Layer Normalization 1: (None, 4, 120) — Further normalizes the processed sequences to ensure consistency and stability.

A sequence pooling layer extracts the most informative features from the encoded patches, which are then fed into a fully connected dense layer for final classification. The DRCCT model, with 2,342,326 parameters, is trained using the Adam optimizer over 100 epochs, demonstrating robust learning and excellent generalization capabilities. This comprehensive architecture enables the model to effectively analyze retinal images and classify diabetic retinopathy severity with high accuracy.

4. Results and Discussion

4.1. Training and Validation Performance

Training and Validation: The performance of the DRCCT model in terms of training and validation accuracy was outstanding over 100 epochs, completed in approximately 423 seconds. Throughout the training period, the validation accuracy consistently surpassed the training accuracy, suggesting that the model did not overfit the training data. This observation highlights the model's ability to generalize and adapt to new, previously unseen data. Figure 9 and 10 demonstrates the effectiveness of the DRCCT model in learning from training data and generalizing well our model.



```
·· Model: "model"
   __________________________________________________________________________________________________
    Layer (type)                Output Shape                 Param #    Connected to
   ==================================================================================================
    input_1 (InputLayer)        [(None, 112, 112, 3)]        0          []

    cct_tokenizer (CCTTokenize  (None, 4, 128)               97200      ['input_1[0][0]']
    r)

    tf.__operators__.add (TFOp  (None, 4, 128)               0          ['cct_tokenizer[0][0]']
    Lambda)

    layer_normalization (Layer  (None, 4, 128)               256        ['tf.__operators__.add[0][0]']
    Normalization)

    multi_head_attention (Mult  (None, 4, 128)               527488     ['layer_normalization[0][0]',
    iHeadAttention)                                                      'layer_normalization[0][0]']

    stochastic_depth (Stochast  (None, 4, 128)               0          ['multi_head_attention[0][0]']
    icDepth)

    add (Add)                   (None, 4, 128)               0          ['stochastic_depth[0][0]',
                                                                         'tf.__operators__.add[0][0]']

    layer_normalization_1 (Lay  (None, 4, 128)               256        ['add[0][0]']
   ...
   Total params: 2342326 (8.94 MB)
   Trainable params: 2342326 (8.94 MB)
   Non-trainable params: 0 (0.00 Byte)
```

Figure 9. Model Composition.

4.2. Model Testing and Metrics

The following table presents the results for the entire APTOS dataset, with an 80% split for training and 20% for validation. We use accuracy and precision, defined by Equations (8) and (9) respectively, as metrics to assess the correctness of our classification model.

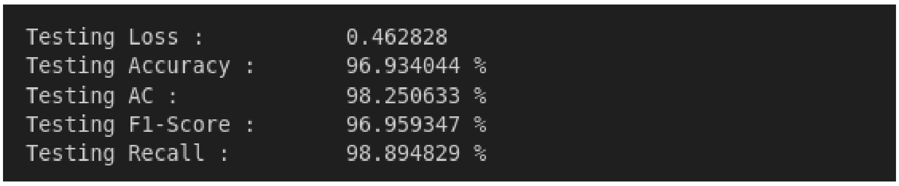
$$\boldsymbol{Accuracy} = (\boldsymbol{TP} + \boldsymbol{TN})/(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}) \qquad (8)$$

$$\boldsymbol{Precision} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FP}) \qquad (9)$$

Accuracy measures the overall effectiveness of the model in correctly classifying both positive and negative instances, while precision focuses on the accuracy of the positive predictions.The F1 score combines precision and sensitivity into a single metric. It is calculated using equation 9.

$$\boldsymbol{F1\ score} = \boldsymbol{2} \times (\boldsymbol{precision}\ \times \boldsymbol{recall})/(\boldsymbol{precision} + \boldsymbol{recall}) \qquad (10)$$

The model was tested on a new dataset to evaluate its generalization performance. It achieved high accuracy, with precision and recall scores of 96.93% and 98.89%, respectively. Various metrics, including sensitivity, specificity, precision, and F1 score, were used to assess its performance. These metrics confirm the model's ability to correctly classify samples across different classes, effectively balancing false positives and negatives.

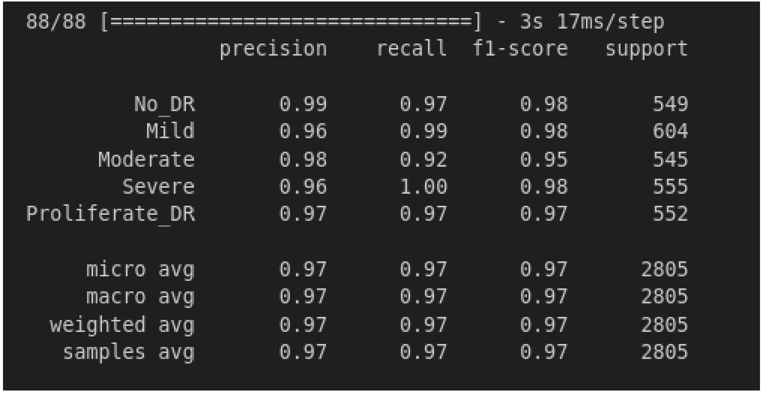


```
Testing Loss :              0.462828
Testing Accuracy :          96.934044 %
Testing AC :                98.250633 %
Testing F1-Score :          96.959347 %
Testing Recall :            98.894829 %
```

Figure 10. Model Performance.

The close alignment between training and validation losses suggests minimal overfitting and robust performance, supporting the model's high precision and recall scores across various diabetic retinopathy stages. This balance between training and validation losses underscores the model's reliability and effectiveness in accurately classifying diabetic retinopathy.

As shown in Figure 11, our results indicate that our model performs well, with an average F1 score of 0.973 across all classes. This confirms that the model is effective at detecting true positives while minimizing false positives.



```
88/88 [==============================] - 3s 17ms/step
                precision    recall  f1-score   support

        No_DR        0.99      0.97      0.98       549
         Mild        0.96      0.99      0.98       604
     Moderate        0.98      0.92      0.95       545
       Severe        0.96      1.00      0.98       555
Proliferate_DR       0.97      0.97      0.97       552

    micro avg        0.97      0.97      0.97      2805
    macro avg        0.97      0.97      0.97      2805
 weighted avg        0.97      0.97      0.97      2805
  samples avg        0.97      0.97      0.97      2805
```

Figure 11. Multi-Class Classification Performance of the Model: Metrics and Results.

4.3. Confusion Matrix Analysis

These metrics demonstrate that the model can balance the trade-off between avoiding false positives and false negatives and can capture most relevant samples for each class. Confusion Matrix: The confusion matrix presented in Figure 12 evaluates the performance of the classification model for diagnosing diabetic retinopathy.

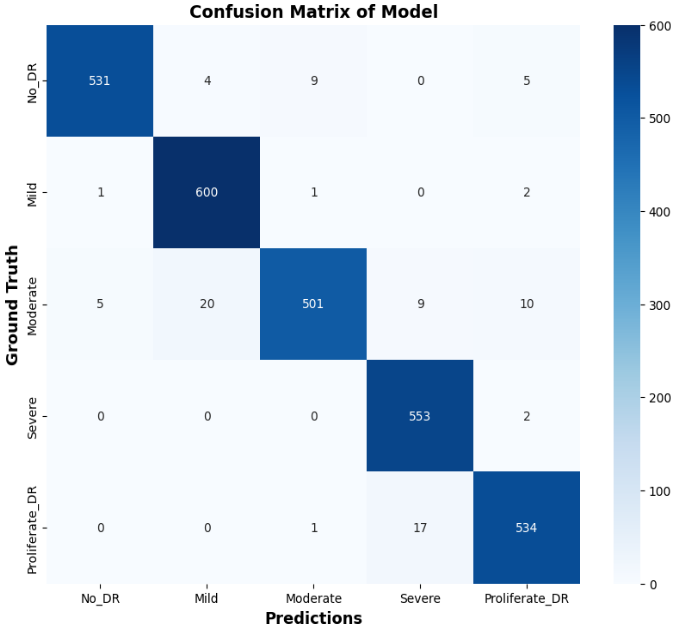


Figure 12. Confusion Metric of DRCCT.

The confusion matrix reveals that while the DRCCT model demonstrates a strong performance in certain classes like No_DR, it struggles significantly with others, particularly Mild, Moderate, Severe, and Proliferate_DR. The high number of false positives in the Mild class and false negatives in the Moderate class suggest that the model may require further fine-tuning or additional data to better distinguish between these stages. The complete miss in the Severe and Proliferate_DR classes indicates a need for model refinement, especially considering the clinical importance of accurately detecting these advanced stages of diabetic retinopathy.

- **True Positives (TP):** Correctly identified positive cases.
- **True Negatives (TN):** Correctly identified negative cases.

doi:10.20944/preprints202409.0303.v1

- **False Positives (FP):** Incorrectly identified positive cases.
- **False Negatives (FN):** Incorrectly identified negative cases.

Specificity as defined by Equation 11, is sometimes known as the false positive rate (FPR). It is the counterpart to sensitivity and measures the model's ability to correctly identify negative samples. Mathematically, specificity is expressed as follows:

$$\textbf{Specificity} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (11)$$

Precision, also referred to as positive predictive value (PPV), assesses the ratio of correctly predicted positive results to the total number of predicted positives. It is given by Equation 12:

$$\textbf{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (12)$$

4.4. Training and Validation Loss

The training and validation loss values, as depicted in Figure 2, demonstrate the DRCCT model's effective learning and generalization. The consistently low training loss, decreasing to around 0.03, indicates that the model is effectively minimizing errors on the training dataset. Similarly, the validation loss, stabilizing at approximately 0.04, reflects the model's strong ability to generalize to new, unseen data.

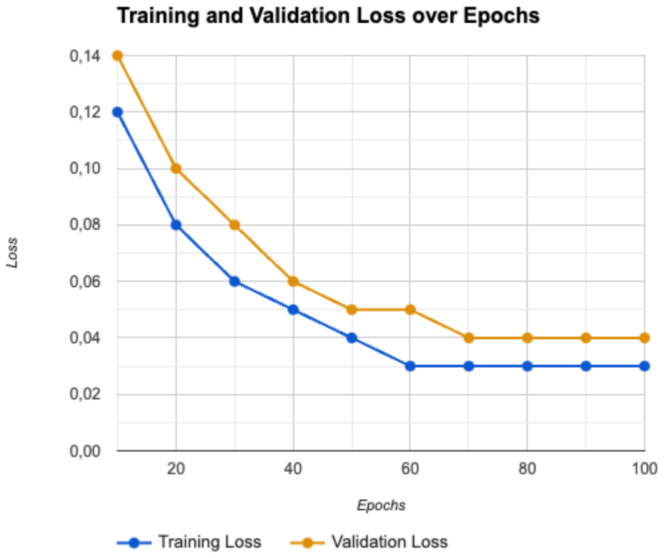


Figure 13. Training and validation loss.

4.5. Advanced Optimization Strategies

4.5.1. Optimizer

In the DRCCT model, the AdamW optimizer was selected for its superior handling of weight decay, which effectively prevents overfitting by decoupling weight decay from the gradient update process. AdamW's adaptive learning rate mechanism adjusts for each parameter based on gradient moments, ensuring stable and efficient convergence. Its momentum-based updates and bias correction techniques further enhance training stability and speed, making it well-suited for the model's complex architecture, which combines convolutional networks and transformers.

The AdamW update rule is given by:

$$\theta_{\mathrm{t}} = \theta_{\mathrm{t-1}} - \eta(\frac{\widehat{\mathrm{m}}_{\mathrm{t}}}{\sqrt{\hat{\mathrm{v}}_{\mathrm{t}}}+\epsilon} + \lambda \times \theta_{\mathrm{t-1}}) \quad (13)$$

- θt represents the parameters at time step t,
- η is the learning rate,
- $\widehat{\mathrm{m}}_{\mathrm{t}}$and $\hat{\mathrm{v}}_{\mathrm{t}}$ are the bias-corrected first and second moment estimates,

doi:10.20944/preprints202409.0303.v1

- $\theta \quad = 10^{-8}$ is a small constant for numerical stability,
- λ= 0.01 is the weight decay coefficient.

4.5.2. Cost Function

Loss functions are crucial in guiding a model's learning process by minimizing errors and improving performance. The selection of a loss function depends on the specific goals and data characteristics. For the DRCCT model, Categorical Cross-Entropy was chosen due to its effectiveness in multi-class classification tasks, such as diabetic retinopathy severity classification. It measures the divergence between the model's predicted probability distribution and the actual distribution of the classes, directly aiding in accuracy improvement. The Categorical Cross-Entropy loss function is defined as:

$$Loss_{CEE} = -\sum_{n=1}^{N} (y_i \log(\hat{y}_i)) \quad (14)$$

where:

- N is the number of classes,
- y_i is the ground truth label for class i,
- $\hat{y}_i$ i is the predicted probability for class i.

In cases where data imbalance is a concern, Focal Loss can be an alternative, focusing more on harder-to-classify examples and ensuring balanced performance across all classes. The Focal Loss function is given by:

$$Loss_{CEE} = -\sum_{n=1}^{N} (1-\hat{y}_i)^{\gamma}(y_i \log(\hat{y}_i)) \quad (8)$$

Where:

- γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted.

4.5.3. Learning Rate Adjustment

The learning rate is one of the most crucial hyperparameters in deep learning, governing how quickly or slowly a model adapts to the training data. Selecting an appropriate learning rate is essential for effective convergence. In this work, several techniques were implemented to optimize the learning rate:

Cyclical Learning Rate (CLR): CLR varies the learning rate cyclically between a minimum and maximum value. By allowing the learning rate to periodically increase and decrease, the model can escape local minima, leading to better convergence. We set the base learning rate at 1e-4 and the maximum at 1e-3, which helped in stabilizing the training process and achieving more robust performance.

4.5.4. Regularization Techniques

Regularization is essential to prevent overfitting, especially when dealing with a large number of parameters, as is common in transformer-based models like DRCCT. Several regularization methods were utilized:

Dropout: To mitigate overfitting, dropout was increased in the Transformer blocks from 0.3 to 0.4. By randomly deactivating a fraction of neurons during training, dropout forces the model to learn more robust features that are not reliant on specific neurons.

L2 Regularization: Also known as weight decay, L2 regularization penalizes large weights by adding a regularization term to the loss function. This prevents the model from becoming overly complex and helps in maintaining generalization. In this study, an L2 regularization coefficient between 1e-4 and 5e-4 was applied.

Label Smoothing: To further reduce overfitting and make the model more tolerant to noisy labels, label smoothing was employed with a smoothing factor of 0.1. This technique reduces the confidence of the model in its predictions, which can help in preventing the model from becoming too confident and overfitting to the training data.

4.6. Results Overview

The DRCCT model shows consistently high performance across all metrics, effectively classifying diabetic retinopathy (DR) into five categories: No_DR, Mild, Moderate, Severe, and Proliferate_DR. Our model for classifying diabetic retinopathy (DR) across different severity levels demonstrates consistently high performance, with impressive metrics throughout.

Table 3 reveals that the model demonstrates exceptional precision and recall across all diabetic retinopathy (DR) classes. Notably, it achieves outstanding performance in the No_DR category, with a precision of 0.99, recall of 0.97, and an F1-score of 0.98. Similarly, in the Severe DR category, the model exhibits a precision of 0.96, a perfect recall of 1.00, and an F1-score of 0.98. These results underscore the model's strong capability in accurately identifying both the absence and presence of severe DR. This high level of accuracy and the model's effectiveness in capturing true positives are crucial in medical diagnostics, ensuring reliable identification of DR stages. Additionally, the F1-scores, ranging between 0.95 and 0.98 across all classes, reflect the model's ability to balance precision and recall, making it dependable across various stages of DR. The DRCCT model stands out with its balanced performance across all DR classes, minimal overfitting, and high accuracy metrics. It is competitive with the latest models in the field and shows potential for practical application in medical diagnostics, where reliability and accuracy are critical.

Table 3. Comprehensive Results Summary Table.

Metri	No_DR	Mild	Moderate	Severe	Proliferate	Micro Avg	Macro Avg	Weighted Avg	Samples Avg
Recall	0.99	0.96	0.98	0.96	0.97	0.97	0.97	0.97	0.97
Recall	0.97	0.99	0.92	1.00	0.97	0.97	0.97	0.97	0.97
F1-Score	0.98	0.98	0.95	0.98	0.97	0.97	0.97	0.97	0.97
Support	549	604	545	555	552	2805	2805	2805	2805

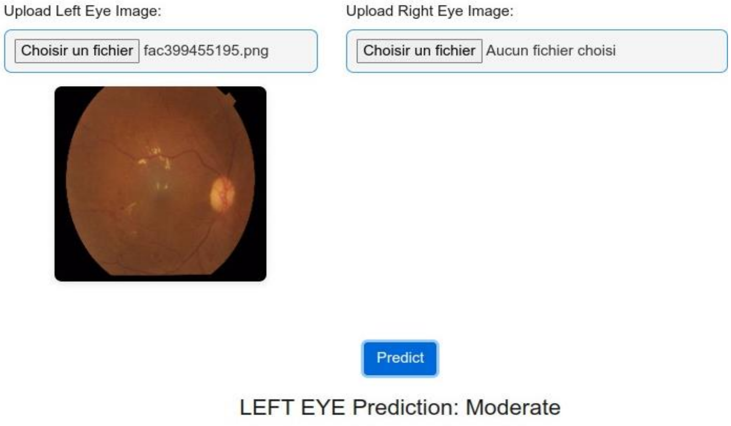
4.7. Comparative Study of Results

As demonstrated in Table 4, The DRCCT (our model) exhibits outstanding performance across several key metrics, including precision, recall, and F1-score, all achieving a value of 0.973 across all categories. This performance is not only significantly higher than many of the other models listed in related work but also shows consistent accuracy and recall across different diabetic retinopathy (DR) severity levels. Such consistency is particularly notable when compared to models like the ResNet-50 with Random Forest classifier and ViT CNN, which exhibit varying performance across different datasets. This highlights the robustness and reliability of our model for real-world applications. Additionally, while models like the Xception pretrained model and Residual Block + CNN report high accuracy, they often lack detailed information on precision, recall, or F1-score across multiple classes. In contrast, our model provides comprehensive metrics that clearly indicate its balanced performance across all severity levels of DR. The advanced architecture of DRCCT leverages the strengths of both convolutional networks and transformers, particularly through the use of a Compact Convolutional Transformer approach. This allows the model to more effectively capture spatial features and relationships within retinal images, which is crucial for accurate DR classification.When compared to state-of-the-art systems like the Residual-Dense System and Vision Transformers with Masked Autoencoders (MAE), our model's F1-score is on par or better, showcasing its ability to compete with more complex architectures while maintaining efficiency. Overall, DRCCT (2024) offers balanced and superior performance, making it a strong candidate for deployment in clinical settings. It not only excels in key metrics but also demonstrates the robustness needed for reliable DR severity classification, solidifying its position as an advanced and practical solution for diabetic retinopathy diagnosis.

Table 4. This is a table. Tables should be placed in the main text near to the first time they are cited.

Authors	Method	Performance	Our Model
Sheikh, S., & Qidwai, (2020)	Transfer Learning of MobileNetV2	90.8% DR, 92.3% RDR	Likely superior, F1-score: 0.97
Gao et al., (2019)	DL/Efficient CNN	90.5% Accuracy	better
Yaakoob et al. (2021)	ResNet-50 with a Random Forest classifier	96% on the Messidor-2 ,75.09% EyePACS	Better than EyePACS, comparable to Messidor-2
Dharmana, M al., (2020)	Blob Technique and Naïve Bayes	83% Accuracy	Significantly better
Wang, J & al., 2020	Deep Convolutional Neural Networks	Kappa 0.8083	Likely superior based on F1-score
Toledo-Cortés et al., (2020)	Deep Learning/DLGP-DR, Inception-V3	93.23% Sensitivity, 91.73% Specificity, 0.9769 AUC	Comparable, slightly better F1-score
Wang, S. & al., (2020)	Deep Learning/GAN Discriminative model	EyePACS: 86.13% Accuracy, Messidor: 84.23% Accuracy, Messidor(2): 80.46% Accuracy	Superior performance across metrics
Touati, M., Nana, L., Benzarti, F. (2023)	Xception pretrained model	Training accuracy: 94%, Test accuracy: 89%, F1 Score: 0.94	Better F1-score: 0.97
Toledo-Cortés et al., 2020	Deep Learning/ DLGP-DR, Inception-V3	93.23% Sensitivity, 91.73% Specificity, 0.9769 AUC	Better in Sensiftiviy and Specificiy
Z. Wang, Y. Yin. (2017)	Deep Learning/CNN+Attention Network	AUC 0.921 Acc 0.905 for normal/abnormal	Likely superior based on metrics
Khan, I et al. (2023)	Compact Convolution Network	Acc 90.17%	Significantly better, likely 97% accuracy
M. BERBaR (2022)	Residual-Dense System	97% in classifying DR severity	Comparable or slightly better
Nazi et a (2023)	ViT CNN	F1-score: 0.825, accuracy: 0.825, B Acc: 0.826, AUC: 0.964, precision: 0.825, recall: 0.825, specificity: 0.956.	Significantly better, F1-score: 0.97
Ijaz Bashir, et al. (2023)	Residual Block + CNN	Accuracy of 97.5%	Comparable, accuracy likely around 97%
Yasashvini R et al. (2022)	hybrid CNNs ResNet, and a hybrid CNN with DenseNet	.The models achieved accuracy rates of 96.22%, 93.18%, and 75.61%, respectively	Better
Yaoming Yang et al. (2024)	Vision Transformers (ViT) combined with Masked Autoencoders (MAE)	accuracy 93.42% ,AUC 0.9853, sensitivity 0.973, specificity 0.9539	Slightly better F1-score

As a researcher in LabSTICC and Pixemantic Startup, specializing in deep healthcare, a create cutting-edge platformis created for diabetic retinopathy detection. Pixemantic Develeopors helped deploy a web application powered by our DRCCT model, which has shown remarkable accuracy, including in challenging left eye cases, as seen in Figure 14. This project, supported by Dr. Rabeb Touati's expertise exemplifies the power of AI in advancing early diagnosis and improving patient care.



Upload Left Eye Image:

Choisir un fichier fac399455195.png

Upload Right Eye Image:

Choisir un fichier Aucun fichier choisi

LEFT EYE Prediction: Moderate

Figure 14. Left Eye Prediction.

5. Conclusion

This research has effectively developed the Diabetic Retinopathy Compact Convolutional Transformer (DRCCT) model, demonstrating its high precision in classifying and detecting the stages of diabetic retinopathy. By merging convolutional layers with transformer techniques, the DRCCT model achieved remarkable results, including an average F1 score of 0.973, a precision of 96.93%, and a recall of 98.89%. The model's strong performance, without overfitting and high validation accuracy across 100 training epochs, highlights its ability to accurately identify and differentiate between the various stages of diabetic retinopathy. It surpasses existing models such as MobileNetV2, ResNet-50 with Random Forest classifiers, and Vision Transformers with Masked Autoencoders, offering superior precision and robustness in addressing class imbalance and reducing false positives. The successful use of advanced regularization techniques, such as dropout and stochastic depth, emphasizes the model's versatility and its potential for clinical integration, where early and accurate detection is vital for effective treatment.

6. Future Research

Integrating watermarking techniques into the DRCCT model marks a significant step forward in enhancing data security and integrity within medical diagnostics. By embedding watermarks using an encoder-decoder framework, diagnostic outputs can be effectively safeguarded against tampering and unauthorized access. This method is particularly crucial in telemedicine, where medical images and diagnostic data are often transmitted over insecure networks. Watermarking not only protects the data but also ensures its authenticity, thus preserving trust in AI-assisted diagnoses and protecting sensitive medical information. Future research should aim to refine watermarking methods to increase their robustness and contextual relevance. Creating watermarks based on key diabetic retinopathy features will ensure that these security measures are both effective and tailored to the medical setting. Additionally, investigating predictive watermarking techniques that adapt to changing disease states could further enhance data integrity and reliability, offering a dynamic and secure framework for managing patient information and improving diagnostic accuracy.

Author Contributions: For the research article, the following contributions were made by the authors: **Conceptualization**: Mohamed Touati and Rabeb Touati. **Methodology**: Mohamed Touati. **Software**: Mohamed Touati. **Validation**: Laurent Nana, Rabeb Touati, and Faouzi Benzarti. **Formal Analysis:** Mohamed Touati.

Investigation: Mohamed Touati and Rabeb Touati. **Resources**: Faouzi Benzarti. **Data Curation**: Mohamed Touati **Writing—Original Draft Preparation**: Mohamed Touati. **Writing—Review & Editing**: Rabeb Touati and Laurent Nana. **Visualization**: Mohamed Touati and Sadok Ben Yahia **Supervision**: Faouzi Benzarti Laurent Nana, Rabeb Touati and Sadok Ben Yahia **Project Administration**: Laurent Nana. **Funding Acquisition**: Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable, as the study did not involve humans or animals.

Informed Consent Statement: Not applicable, as the study did not involve humans.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: We acknowledge the support from the University of Western Brittany (UBO), Lab-STICC, University of Tunis, and the Faculty of Medicine of Tunis. Special thanks to all technical staff involved in this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Additional Perspective: As part of our future work, we plan to integrate a watermarking technique into the encoder side of our model. This watermarking method will serve as a security feature to protect the integrity of the data and enhance cryptographic measures within the encoder-decoder framework. The development of this method aims to ensure secure and reliable transmission and storage of sensitive medical data.

References

1. Bidwai, P.; Gite, S.; Pahuja, K.; Kotecha, K. A Systematic Literature Review on Diabetic Retinopathy Using an Artificial Intelligence Approach. *Big Data Cogn. Comput.* **2022**, *6*, 152. https://doi.org/10.3390/bdcc6040152
2. Subramanian, S.; Mishra, S.; Patil, S.; Shaw, K.; Aghajari, E. Machine Learning Styles for Diabetic Retinopathy Detection: A Review and Bibliometric Analysis. *Big Data Cogn. Comput.* **2022**, *6*, 154. https://doi.org/10.3390/bdcc6040154
3. S. et al. « Transformers in medical imaging: A survey". In: Medical Image Analysis. » (2023), adresse : https://doi.org/10.1016/j.media.2023.102802.
4. Sheikh, S., & Qidwai, U. (2020). Using MobileNetV2 to Classify the Severity of Diabetic Retinopathy. International Journal of Simulation- -Systems, Science & Technology, 21(2).
5. J. Gao, C. Leung and C. Miao, "Diabetic Retinopathy Classification Using an Efficient Convolutional Neural Network", in 2019 IEEE InternationalConferenceonAgents (ICA),2019
6. JWang,S.,Wang,X.,Hu,Y.,Shen,Y.,Yang,Z.,Gan,M.,&Lei,B. (2020). Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision. IEEE Transactions on Automation Science and Engineering, 18(2), 574-585
7. Touati, M., Nana, L., Benzarti, F. (2023). A Deep Learning Model for Diabetic Retinopathy Classification. In: Motahhir, S., Bossoufi, B. (eds) Digital Technologies and Applications. ICDTA 2023. Lecture Notes in Networks and Systems, vol 669. Springer, Cham.
8. Yaqoob, M. K., Ali, S. F., Bilal, M., Hanif, M. S., & Al-Saggaf, U. M. (2021). ResNet based deep features and random forest classifier for diabetic retinopathy detection. Sensors, 21(11), 3883.
9. Dharmana, M. M., & Aiswarya, M. S. (2020, July). Pre-diagnosis of Diabetic Retinopathy using Blob Detection. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 98-101). IEEE.
10. Toledo-Cortés, S., De La Pava, M., Perdomo, O., & González, F. A. (2020, October). Hybrid Deep Learning Gaussian Process for Diabetic Retinopathy Diagnosis and Uncertainty Quantification. In *International Workshop on Ophthalmic Medical Image Analysis* (pp. 206- 215). Springer, Cham.
11. Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., & Wang, X. (2017). Zoom- in-net: Deep mining lesions for diabetic retinopathy detection. In Medical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20 (pp. 267- 275). Springer International Publishing.
12. Vo, H. H., & Verma, A. (2016, December). New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In 2016 IEEE International Symposium on Multimedia (ISM) (pp. 209-215). IEEE.

13. Touati, M., Nana, L., & Benzarti, F. (2024). Enhancing diabetic retinopathy classification: A fusion of ResNet50 with attention mechanism. In 10th International Conference on Control, Decision and Information Technologies (CoDIT).
14. Vani, K. S., Praneeth, P., Kommareddy, V., Kumar, P. R., Sarath, M., Hussain, S., & Ravikiran, P. (2024). An Enhancing Diabetic Retinopathy Classification and Segmentation based on TaNet. *Nano Biomedicine & Engineering*, *16*(1).
15. Alwakid, G., Gouda, W., Humayun, M., & Jhanjhi, N. Z. (2023). Enhancing diabetic retinopathy classification using deep learning. *Digital Health*, *9*, 20552076231203676.
16. Al-Hammuri, K., Gebali, F., Kanan, A., & Chelvan, I. T. (2023). Vision transformer architecture and applications in digital health: a tutorial and survey. *Visual computing for industry, biomedicine, and art*, *6*(1), 14.
17. Nazih, W., Aseeri, A. O., Atallah, O. Y., & El-Sappagh, S. (2023). Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access*, *11*, 117546-117561.
18. Khan, I. U., Raiaan, M. A. K., Fatema, K., Azam, S., Rashid, R. U., Mukta, S. H., ... & De Boer, F. (2023). A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time. *Biomedicines*, *11*(6), 1566.
19. M. BERBaR. « Features extraction using encoded local binary pattern for detection and grading diabetic retinopathy. » (2022), adresse: https://doi.org/10.1007/s13755-022-00181-z.
20. Bashir, I., Sajid, M. Z., Kalsoom, R., Ali Khan, N., Qureshi, I., Abbas, F., & Abbas, Q. (2023). RDS-DR: An Improved Deep Learning Model for Classifying Severity Levels of Diabetic Retinopathy. *Diagnostics*, *13*(19), 3116.
21. Y. R. et al. « Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks. » (2022), adresse: https://doi.org/10.3390/sym14091932.
22. N. G. et al. « Evaluation of artifcial intelligence techniques in disease diagnosis and prediction. » (2023), adresse: https://link.springer.com/article/10.1007/s44163-023-00049-5?fromPaywallRec=true.
23. Y. Y. et al. « Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. » (2023), adresse : https://doi.org/10.1371/ journal.pone.0299265.
24. N. S. R. KaRthikEyan. « Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures. » (2021), adresse: https://ieeexplore.ieee.org/abstract/ document/9651024.
25. J. W. et al. « Vision Transformer-based recognition of diabetic retinopathy grade. » (2021), adresse: https://doi.org/10.1002/mp.15312.
26. A. D. et al. « An image is worth 16x16 words: Transformers for image recognition at scale. » (2020), adresse: https://doi.org/10.48550/arXiv.2010.11929.
27. A.-H. et al. « Vision transformer architecture and applications in digital health: a tutorial and survey. » (2023), adresse: https://doi.org/10.1186/s42492-023-00140-9.
28. I. et al. « "Recent advances in vision transformer: A survey and outlook of recent work. » (2022), adresse: https://doi.org/10.48550/arXiv.2203.01536.
29. H. et al. « Escaping the big data paradigm with compact transformers. » (2021), adresse : https://www.researchgate.net/publication/350834450_Escaping_the_Big_Data_Paradigm_with_Compact_Transformers.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.