Journal of Computer Science

Original Research Paper

Transforming Retinal Diagnostics: Advanced Detection of Diabetic Retinopathy Using Vision Transformers and Capsule Networks

[1]Vishal Sharma, [1]Rishu, [1]Vinay Kukreja, [1]Ayush Dogra and [2,3]Bhawna Goyal

[1]Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[2]Marwadi University Research Centre Derailment of Engineering, Rajkot, Gujarat, India
[3]Faculty of Engineering, Sohar University, Sohar, Oman

Article history
Received: 25-09-2024
Revised: 25-10-2024
Accepted: 12-11-2024

Corresponding Author:
Vinay Kukreja
Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Email: onlyvinaykukreja@gmail.com

Abstract: Diabetic Retinopathy (DR), nowadays is one of the leading causes of blindness worldwide, it is a severe complication of diabetes mellitus that affects the retina blood vessels. Accurate diagnosis depends on early detection of DR. The study aims to develop a hybrid model that is the combination of a Vision Transformer and Capsule Network (ViT-CapsNet) to classify the DR at early stages. The ViT-CapsNet model is proposed to detect the DR from the retinal images at the early stage. The eyepieces public dataset is used. The data preprocessing takes place in which the resizing and data augmentation are used to improve the quality and increase the diversity of the data. Then, the Vision transformer extracts the global features from the retinal fundus image while the capsule network preserves the spatial relationships and hierarchies within the data, also classified into different classes that are No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. The ViT-CapsNet model has a precision, recall and F1-Score with values of 0.92, 0.91 and 0.91 respectively. The ViT-CapsNet model shows an accuracy of 94% compared to the other traditional methods such as CNN (88%), ResNet (90%), and EfficientNet (92%). The AUC-ROC scores for classes No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR are 0.56, 0.48, 0.44, 0.45, and 0.51 respectively.

Keywords: Vision Transformers, Capsule Networks, Diabetic Retinopathy, Retinal Images, Deep Learning

Introduction

Diabetes is a fatal disease that is caused by impairment in the beta cells of the pancreas. This leads to alteration in the production of insulin, causing hyperglycemia of more than 120 mg/dL. It can be further categorized into two types type I diabetes and type II diabetes. In type I diabetes there is destruction of beta cells of islet of Langerhans causing insufficient insulin production. Contrarily, Type II diabetes is a disorder where the body produces insulin but the receptors do not react to it (Dilmurodovna, 2023). The contributing reasons for this include Hyperglycemia, Oxidative Stress, Inflammation, and genetic factors. These further result in a decrease in growth factors (vasoendothelial growth factor and platelet-derived growth factor), nucleic acids, and proteins (ElSayed *et al*., 2023). In addition to this, there is also an alteration in the pathways which include the hexosamine pathway, polyol pathway, Nuclear Factor kappa-light-chain-enhancer of activated B cells (NFkB) pathway, Insulin signaling Pathway, Adenosine Monophosphate (AMP) Activated Protein Kinase, Peroxisome Proliferator-Activated Receptor (PPAR) Pathway, Toll-Like Receptor (TLR) Pathway, Gluconeogenesis Pathway, Glycolysis Pathway and Tricarboxylic Acid (TCA) Cycle (Roglic, 2016). The global prevalence of diabetes is around 500 million. Moreover, if diabetes is not controlled within a shorter period it leads to various complications. DR is one of them having a global prevalence of an estimated 27.0% of diabetes patients worldwide, resulting in 0.4 million blind people worldwide. According to a pooled study of many hospital-based studies, 19.48% of Ethiopians and 31.6% of Africans are estimated to have DR.

DR is a condition in which there is damage in the retina, leading to microaneurysms, retinal swelling, and neovascularization, which can cause vision impairment or blindness (Nazih *et al*., 2023). The disease is triggered by hyperglycemia, oxidative stress, inflammation, and dysregulated growth factors, which exacerbate the damage to retinal vessels. Changes in blood flow and pressure also contribute to the progression of this disease. DR is categorised into mainly two types Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). The initial stage is the NPDR of DR which is classified into mild, moderate, and severe, and develops on the retinal blood vessel damage. PDR is a long-term diabetes that affects the retinal blood vessels (Mohan *et al*., 2022). Within PDR, neovascularization, or aberrant blood vessel growth, occurs on the surface of the retina and within the vitreous fluid, which stops or leaks retinal blood vessels, reducing the amount of blood that reaches the retina. Since the early stage of DR has no symptoms, individuals may not become aware of any anomalies in their eyesight until the illness has gotten worse phases (Barker *et al*., 2023). At these late stages, retinal damage may become irreversible, leading to outcomes indicating retinal detachment and vitreous bleeding. Medical professionals can take action to stop or slow the disease's growth before these issues arise by using treatments like intravitreal injections of Vascular Endothelial Growth Factor (VEGF) medications or laser photocoagulation. To diagnose DR routine eye exams are essential for detecting mild changes in the retina. These early indicators can be recognized with the aid of methods such as fundus photography, Optical Coherence Tomography (OCT), and fluorescein angiography (Bajwa *et al*., 2023). The chance of DR getting worse can be reduced by prompt management and therapy, such as cholesterol, blood pressure, and glucose optimization. Additionally, early identification preserves autonomy and mental health by preventing blindness and lowering the need for expensive therapies. The automated technologies enhance clinical decision-making, streamline administrative tasks, and increase diagnostic precision. Early disease detection and personalized treatment are made possible by the systems' rapid analysis of medical data through the use of robotics, machine learning, and artificial intelligence (Phillip *et al*., 2023). By automating tedious chores, lowering mistakes, and relieving administrative burdens, they optimize processes. Robotically assisted surgery enhances accuracy and shortens recovery periods. Real-time patient health assessments are made possible by automated monitoring technology, which allows for timely action. The efficiency, patient-centeredness, and accuracy are enhanced by automated systems. On the other hand, streamlining administrative tasks and human errors also reduced.

Problem Statement

DR is a chronic eye disease that causes blindness and significant vision impairment. The complexity of the retinal images and the subtlety of drug resistance anomalies present formidable challenges to existing diagnostic techniques (Silberman *et al*., 2010). DR diagnosis is hampered by several major problems, including the subjective nature of manual evaluation, the inconsistent performance of automated diagnostic techniques, and variability in retinal image appearance owing to DR stages and patient variations. The inability of current machine learning methods to reliably collect and analyze pertinent information from retinal pictures might result in variability in diagnostic accuracy and possible delays in recognizing key phases of diabetic retinitis (Skouta *et al*., 2023). These drawbacks highlight the need for better techniques that can handle the variety of DR symptoms and increase the accuracy of diagnosis.

Objective

- Developing a hybrid model ViT-CapsNet for DR detection
- The use of ViT enhances the feature extraction from retinal images which improves the DR detection
- Integration with capsule network classifies the different classes of DR

Literature Review

DR is recognized as one of the leading causes of blindness worldwide. The early identification of DR helps shield the patient from the dangers of blindness. The identification and grading of DR were addressed by several methods that used manually created features, such as blood vessels, hemorrhages, micro-aneurysms, and exudates.

Traditional Methods

For many years, the diagnosis of DR has been made using traditional procedures that combine sophisticated imaging tools with a manual examination. The authors suggest that direct ophthalmoscopy can be used as a diagnostic tool for family physicians to screen for DR with results on groups A, B, and O found to have corresponding sensitivity values of 59, 51, and 78% (Nourinia *et al*., 2023). Then comparing the ability of indirect ophthalmoscopy, B-Scan ultrasonography, and ultra-widefield fundus imaging to identify retinal fractures in cataractous eyes with results postoperative Intraocular Diameter Observation (IDO) findings revealed that Ultrasonography (USG (100%)) and preoperative IDO (99%) were the index tests with the highest sensitivity and specificity (Miao *et al*., 2024). The deep learning algorithms to identify DR in retinal fundus photos through a meta-analysis and systematic review with having research design Preferred Reporting Items for

Systematic Reviews and Meta-Analyses proposed by the author which has a sensitivity of 0.83 and specificity of 0.92 (Islam *et al.*, 2020). The fundus fluorescein angiography images of patients with DR could be automatically interpreted and clinically evaluated using deep learning with an accuracy of 91% (Gao *et al.*, 2023). In eyes with referable nonproliferative DR, deep capillary nonperfusion on OCT angiography predicts complications, according to the author's theory with results having a sensitivity of 89% and specificity of 98% (Ong *et al.*, 2023). The superficial Optical Coherence Tomography Angiography (OCTA) demonstrated the best performance in deep learning classification of DR, with 87.25% accuracy, 78.26% sensitivity, and 90.10% specificity, compared to control, No DR and NPDR layers (Ebrahimi *et al.*, 2023). Automated Retinal Image Analysis for DR Screening in a Primary Care Setting Increases Adherence to Ophthalmic Care, this method has a 100% sensitivity rate for identifying abnormal screening results and, a 65.7% specificity rate (Liu *et al.*, 2021). The authors presented an AI-based smart teleophthalmology application for diagnosing DR, achieving a precision of 94.44%, specificity of 91.35%, and sensitivity of 92.51% (Ghouali *et al.*, 2022). The different types of traditional methods are shown in Fig. (1).

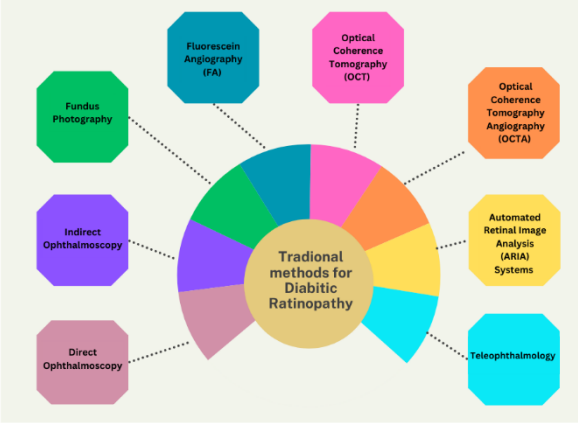


Fig. 1: Different types of traditional methods

Deep Learning (DL) and Machine Learning (ML) Methods

Many DL and ML techniques have been used for DR identification and each model has particular advantages. The author proposed a random forest algorithm to diagnose DR with a sensitivity of 91.40% and an accuracy of 94.38% (Zaaboub and Douik, 2020). Using the datasets IDRiD and DIARETDB1, a multidomain bioinspired feature extraction and selection model with 96.5% accuracy in DR severity level identification (Uppamma and Bhattacharya, 2023). Generative Adversarial Network (GAN), a deep neural network with a discriminator and a generator has shown remarkable results in image synthesis and image-to-image translation. Image domains are utilized in ophthalmology for segmentation, super-resolution, post-intervention prediction, data augmentation, denoising, and feature extraction, but have limitations like mode collapse and spatial deformities (You *et al.*, 2022). The other techniques used are shown in Table (1).

Challenges

Traditional methods, while highly interpretable, are limited by their reliance on manual feature extraction, limited scalability, and poor generalization to diverse datasets, resulting in inconsistent performance and a lack of adaptability to new data. Then to overcome these challenges ML and DL techniques take place but they also have limitations (Sebastian *et al.*, 2023). In ML, the DL models, despite their improved scalability, adaptability, and performance, often struggle with interpretability and overfitting if not managed carefully. They also face difficulties in dealing with low-quality images, which can be mitigated through advanced augmentation techniques, despite their large amount of labeled data and substantial computational resources (Alyoubi *et al.*, 2020). Table (2) shows the challenges between the traditional and ML models.

Table 1: Relevant studies with their work on datasets with results and techniques used

Reference no.	Techniques	Results			Dataset
Bilal *et al.* (2022)	Convolutional Neural Network (CNN)	Sensitivity 0.94	Specificity 0.96	Accuracy 0.96	EyePACS-1
		Sensitivity 0.93	Specificity 0.94	Accuracy 0.93	Messidor-2
		Sensitivity 0.91	Specificity 0.92	Accuracy 0.91	DIARETDB0
Al-hazaimeh *et al.* (2022)	Deep Convolutional Neural Network (DCNN)	98.8			Eye-PACS
Franklin and Rajan (2013)	Artificial Neural	0.953			DRIVE

	Network (ANN)					
Oh *et al*. (2021)	Residual Network with 34 Layers (ResNet-34)	Performance metric ETDRS 7SF ETDRS F1–F2	Accuracy 0.82 0.80	Sensitivity 0.82 0.80	Specificity 0.82 0.80	Catholic Kwandong University International St. Mary's Hospital provided the UWF fundus photos.
Islam *et al*. (2023)	Swin Transformer and Residual Network with 152 Layers, Version 2 (ResNet152V2)	Accuracy 0.99045	precision 0.9907	recall 0.9905	F1 score 0.9903	Asia Pacific Tele ophthalmology Society (APTOS)
Hameed Abbood *et al*. (2022)	GANs	94.2%				IDRID and MESSIDOR
Carrera *et al*. (2017)	Support Vector Machine (SVM)	95%				Messidor
Qomariah *et al*. (2019)	CNN and SVM	95.20%				Messidor
Priya and Aruna (2013)	SVM	Sensitivity 96.7%		Specificity 71.4%		DIARETDB0
Bilal *et al*. (2022)	U- Net	94.59% 97.92% 93.52%				EyePACS-1 Messidor-2 DIARETDB0
Das *et al*. (2021)	CNN	Accuracy 98.7%		Precision 97.2%		DIARETDB1
Yaqoob *et al*. (2021)	Random Forest and ResNet	96% 75.09%				Messidor-2 EyePACS

Table 2: Challenges that are in traditional methods and machine learning methods with specifications

Specifications	Traditional methods	Machine learning methods
Feature extraction and representation	labor-intensive and manual, prone to missing small patterns	It is automatic, and picks up intricate patterns, but requires big datasets
Efficiency and scalability	Manual processes lead to limited scalability	Quite scalable, but it needs a lot of processing power
Interpretability and trustworthiness	Easily interpreted and decision-making that makes sense	Uninterpretable data are frequently viewed as "black boxes"
Generalization and robustness	Overfitting can occur when there is poor generalization of new data	Enhanced generalization, but insufficient regularization may cause overfitting.
Handling low-quality images	When noise or defects are present in retinal images, performance suffers greatly.	Handel the imbalance images with advanced techniques
User Acceptance	Due to its transparency and ease of use, it is highly regarded by physicians	Due to trade-offs between interpretability and performance, there is mixed acceptance
Speed and real-time processing	slower and unsuitable for real-time use	Faster and also good for real-time use

Materials and Methods

Model Architecture

For early detection and improved retinal image classification, ViT-CapsNet architecture has been proposed as shown in Fig. (2). ViT transformer extracts the features and the capsule network to classify the DR classes. The Vit works in several stages. First, it divides the image into patches with a fixed size. Then these patches are flattened into vectors; further vectors are mapped into high-dimensional space. After that, positional encoding is added to retain its physical information. By getting the physical information from the positional encoding the layer normalization provides stability to the input and accelerates the process and then multi-head attention concentrates on the different parts of the image simultaneously. Then the capsule network has different stages which are the convolutional layer, primary capsule, and digit capsule. In which the convolutional layer extracts the local features, then the primary capsule captures the complex features and the digit capsule gets the information by dynamic routing. The integration of the ViT-CapsNet increases the model's strength. In the classification stage, the output of the capsule network shows the probability of the input image belonging to a particular DR grade, which is No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR. Before starting this phase, data collection and pre-processing are important steps.

Data Collection

The dataset is collected from Kaggle (https://www.kaggle.com/datasets/andreivann/eyepacs/data), which is a publically available dataset for DR disease as shown in Fig. (3). The eyepacs data contains 30262 high-resolution fundus images and the distribution of the images concerning classes is shown in Table (3). The images were gathered from US primary care facilities. The clinicians assigned a severity grade of 0, 1, 2, 3, and 4 which is NO DR, Mild DR, Moderate DR, severe DR, and Proliferative DR to each image.

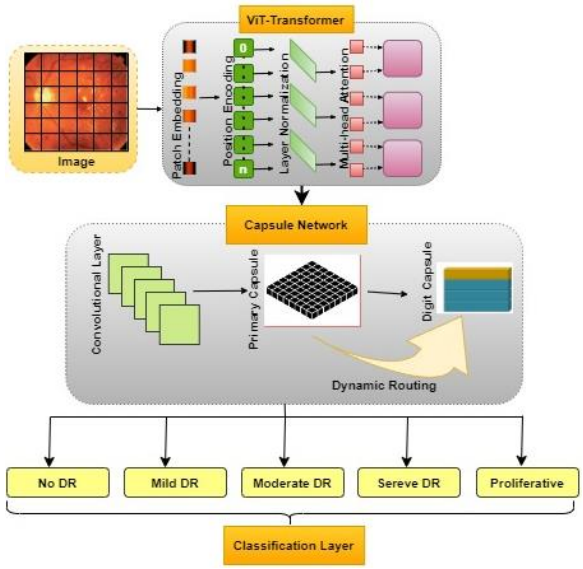


Fig. 2: Vit- transformer and CapsNet

Fig. 3: Eyepacs dataset of retinal images

Table 3: Dataset distribution of eyepieces images

Grade	Classes	Eyepacs images
0	No DR	22,116
1	Mild DR	2,106
2	Moderate DR	4,368
3	Severe DR	845
4	Proliferative DR	827

Each image is authorized by the International Clinical Diabetic Retinopathy (ICDR). The ICDR scale helps healthcare professionals accurately diagnose and monitor the progression of DR in patients. This standardized system allows for consistent communication and treatment planning among medical providers.

Data Pre-Processing

The accuracy and quality of the model depend upon the input images that are used to train the model. That is why the data is pre-processed before being provided to the model and the pre-processing steps are shown in Fig. (4).

Resizing

The quality of any deep learning model may be impacted by the retinal image's fluctuating size. To overcome this issue, the resizing image process has been done with the help of the bicubic interpolation technique. Bicubic interpolation enhances the image quality by considering the surrounding pixel values. To calculate the bicubic interpolation Eq. (1) is used:

$$I(x,y) = \sum_{i=-1}^{2}\sum_{j=-1}^{2} a_{ij} \cdot f(x-i, y-i) \quad (1)$$

where, $I(x, y)$ is the interpolated value at point (x, y), $f(x-i, y-i)$ is the value of the pixel at $(x-i, y-i)$ and a_{ij} is the derived coefficient from the cubic convolutional kernel and determines the weight of the surrounding pixels.

Augmentation

The augmentation process creates the modified versions of the existing data samples by increasing their diversity, it involves color adjustment and geometric alteration. Color adjustment involves brightness, contrast, saturation, and hue. Brightness can be achieved by multiplying the pixel value by the factor. In contrast, it is done by creating the difference between pixel values and the mean values of the image. Then saturation modifies the color appearance in the image. Then Hue changes the color tone of the images and is also used to modify the color appearance.

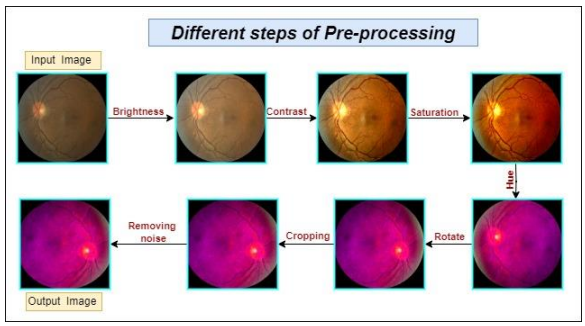


Fig. 4: Pre-processing steps

Table 4: Dataset deviation

Types of datasets	Data deviation (%)
Training dataset	70
Validation dataset	15
Test dataset	15

Geometric alterations are the set of transformations that are used to change spatial properties. It includes rotation, cropping, and removing noise. Rotation rotates the images with a specific angle. Then cropping is involved in selecting the most relevant part that is necessary and removing the unnecessary parts. Then removing noise is used to remove the unwanted pixel variations that are not contributing the information to the image.

Finally, the data is divided into three datasets: Test, validation, and training shown in Table (4). The training dataset has an approximate size of 70%, the validation dataset has an approximate size of 15% and the test dataset has an approximate size of 15%. The model is trained using the training dataset, adjusted and fine-tuned using the validation dataset, and then tested using the test dataset to assess the model's ultimate performance. This process adds to the model's well-trainedness and ability to generalize to new, untested data.

Vit-Transformer

Vit-Transformer is a technique that extracts the features. Features that are local, global, and hierarchical. Its self-attention mechanism captures the contextual informational global dependencies of the image ViT is involved in different stages like patch embedding, position encoding, layer normalization, and multi-head normalization. Firstly, the patch embedding process started.

Patch Embedding

Patch embedding converts the input image into a non-overlapping smaller patch and these patches are embedded into a vector shown in Fig (5). The image size is 1024×11024 pixels which is used for patch embedding shown in Table (5). The purpose of the patch embedding is to convert the high-resolution image into patches and then transform those patches into vectors. The image has three channels (Red, Green, and Blue). To calculate the image patches Eq. (2) is used:

$$N = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right) \quad (2)$$

where, N is the total number of patches, H is the height of the pixel, W is the weight of the pixel and p is the patch size.

Flattening the Patches

It is the process commonly used for converting multi-dimensional images into one-dimensional images and also helps the image to represent the data in a format in which the pixel intensity is treated as a single data point shown in Table (6). To calculate the flattened vector size Eq. (3) is used:

$$Flattened\ vector\ size\ = P \times P \times C \quad (3)$$

where, P is the patch size, C is the channels and the channels are Red, Green, and Blue.

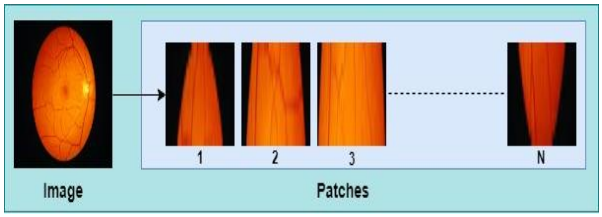


Fig. 5: Patch distribution

Table 5: Number of patches created

Sr no.	Steps	Output
1	Height of the pixel	1024
2	Width of the pixel	1024
3	Patch size	16×16
4	Total number of patches created	$N = \left(\frac{1024}{16}\right) \times \left(\frac{1024}{16}\right) = 64 \times 64 = 4096$

Table 6: Several flattened shapes were created

Stages	Formulas	Outputs
Patch size definition	P × P	= 16×16 = 256
Flatten each patch by which converting each 2D patch matrix into a 1D vector.	$F = Flattened(P) = [P_{11}^R, P_{11}^G, P_{11}^B, \ldots\ldots, P_{16}^R, P_{16}^G, P_{16}^B]$	Every patch turns into a 1D vector with a size of 768
Flattened vector size	= P × P × C	= 16×16×3 = 768
Shape after flattening	Output shape = Total number of vectors × flattened vector size	4096×768

Linear Projection

It is the process that is used to map the flapped images into the lower dimensional space. In ViT image data is converted into a format that the transformer layers can process. It extracts the meaningful features for the classification of the image and captures the relationship and dependencies between different parts of the image shown in Table (7). To calculate the linear projection Eq. (4) is used:

$$E_i = F_i \cdot W + b \quad (4)$$

where, $E_i \in \mathbb{R}^D$ is the projected embedding of the i^{th} patch, $F_i \in \mathbb{R}^d$ is the flattened vector of the i^{th} patch, W $\in \mathbb{R}^{D\times d}$ is the weight matrix, and b $\in \mathbb{R}^D$ is the bias matrix. Here, $\mathbb{R}$ represents the set of real numbers and the real number is used to denote the dimensions of vectors.

Position Embedding

Position encoding is the technique that provides information about the position of each patch. It divides an image into a grid of non-overlapping patches shown in Fig. (6). It provides the spatial location of each patch; spatial location means providing the specific position of each patch with the 2D grid shown in Table (8). Here 2D represents the columns and rows. To calculate the even dimension equation 5 is used:

$$p_{(i,2k)} = sin\left(\frac{i}{10000^{2k/D}}\right) \quad (5)$$

To calculate the odd dimension Eq. (6) is used:

$$p_{(i,2k+1)} = cos\left(\frac{i}{10000^{2k/D}}\right) \quad (6)$$

where, i is the position index, k is the initial dimension index and D is the total embedding dimensions.

Layer Normalization

The layer normalization stabilizes the training process and ensures that the model works and learns properly with a maximum efficiency rate. It is also known as Min-Max normalization because it is a technique that provides a specific range to the data. Min-Max are the original values in the dataset shown in Table (9). Equation (7) is used for layer normalization:

$$x' = \frac{x - min}{max - min} \quad (7)$$

where, x' is the normalization value, The original data's maximum value is denoted by max, the minimum value by min, and x is the original pixel value.

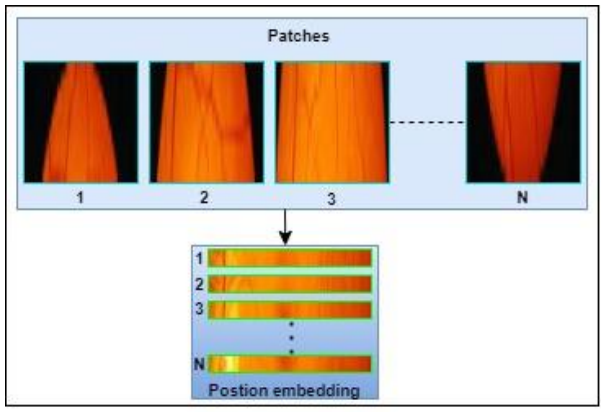


Fig. 6: Position embedding

Table 7: Shapes of the projected image

Steps	Formula	Output
Input flattened vector	$F_i = (f_1, f_2, \cdots f_{768})$	Size of flattened vector 768
Weight matrix	W$\in \mathbb{R}^{768\times D}$	Size of weight matrix = 768× D
Linear transformation	$E_i = F_i \cdot$ W + b	The projected embedded vector size is D
Bias matrix	b $\in \mathbb{R}^D$	It is added to the projection.
Total output shape size after projection	Total number of patches × D	The entire image is represented by 4096× D

Table 8: Position of path embedding

Stages	Formula	Output
Defining parameters	i, k, D	Where i = 10, k = 0, D= 768
Even dimension encoding(2k)	$p_{(i,2k)} = sin\left(\frac{i}{10000^{2k/D}}\right)$	Sin(10)≈ -0.544
Odd dimension encoding(2k+1)	$p_{(i,2k+1)} = cos\left(\frac{i}{10000^{2k/D}}\right)$	Cos($\frac{10}{10000^{2/768}}$)≈ -0.995
Concatenating sin and cos values	$P_i = [P_{(i,0)}, P_{(i,1)}, \cdots, P_{(i,D-1)}]$	$P_{10} = [0.544, 0.995, \cdots, 0.487]$, and here length is 768
Position embedding matrix	$P_i = [P_0, P_1, \cdots, P_{(i,N-1)}]$	$P \in \mathbb{R}^{4096\times 768}$: 4096 rows and 768 columns
Combining with patch embedding	Z = E + P	Combined matrix Z $\mathbb{R}^{4096\times 768}$. Now every patch embedding has a positional context

Table 9: Linear normalization

Stages	Formula	Output
Min- Max and *x* value	Min = 0 Max = 255 X == 128	The minimum value is 0 and the maximum value is 255 for normalization
Linear Normalization	$x' = \frac{x - min}{max - min}$	$x' = \frac{128}{255} \approx 0.502$, the normalized value is in between the range (0,1) and *x* is 128 which is the mid-grayscale value that is 0 to 255

Table 10: Final output size by multi-head attention

Stages	Formulas	Execution	Output
Projected each patch into Q, K, V	$Q = XW^Q$, K= XW^K, V = XW^V	X is the input matrix of size (4096, 256)and project to Q, K, V matrix with shape (4096, 256) where assuming that attention heads are 8 or the size of each head is 64	The shape of each Q, K, V matrix is (4096, 256)
Computing raw attention scores by the dot product of Q and K	Scores = $\frac{Q.K^T}{\sqrt{d_K}}$, where d_K= 64 are the dimensions per head	The raw score for a single head could be [1.2, 0.9, 2.1], and the scaling factor $\sqrt{64}$ = 8	Scaled score is [0.15, 0.11, 0.26]
To obtain attention weights applying softmax function tor to the raw scores	Softmax(Z_i) = $\frac{exp(Z_i)}{\sum_i exp_{(Z_j)}}$	For the scaled scores [0.15, 0.11, 0.26] = softmax(0.15) = $\frac{exp(0.15)}{exp(0.15)+exp(0.11)+exp(0.26)} \approx 0.32$, softmax(0.11) = $\frac{exp(0.11)}{exp(0.15)+exp(0.11)+exp(0.26)} \approx 0.30$, softmax(0.26) = $\frac{exp(0.26)}{exp(0.15)+exp(0.11)+exp(0.26)} \approx 0.38$	Attention weights = (0.32, 0.30, 0.38)
A sum of weighted values	Attention weights × values	Values = [0.5, 1.0, 1.5], Output = (0.32× 0.5, 0.30× 1.0, 0.38× 11.5)	Weighted values = (0.16, 0.30. 0.57)
Concatenation of heads	Concatenating the output from 8 heads, each size is 64	After concatenating the output size = (4096, 512)	Output = (4096, 512)
Final output projection that concatenated back to the input dimension	Concatenated output. W^0, where W^0 is the learnable weight matrix	4096 patches of size 256	Output = (4096, 256)

Multi-head Attention

ViT uses multi-head attention to capture diverse features from input data. Each head learns three linear projections: Queries (Q), Keys (K), and Values (V). The attention mechanism computes compatibility, scales, derives attention weights, and aggregates values. This process enables the model to focus on input parts simultaneously, capturing detailed patterns and relationships shown in Table (10) and to calculate multi-head attention Eq. (8) is used:

$$Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_k}}\right) V_i \quad (8)$$

where, Q_i is the query matrix at the i^{th} head, K_i is the key at the i^{th} head V_i is the value at *the* i^{th} head d_k are the dimensions of the key vectors and softmax converts raw vector scores into probabilities. In which $Q_i = XW_i^Q$:

$$K_i = XW_i^K, V_i = X\,W_i^V$$

where: W_i^Q is the weight matrix of the query for the i^{th} attention head W_i^K is the weight matrix of the key for the i^{th} attention head, W_i^V is the weight matrix of the value for the i^{th} attention head and *X* is the input matrix.

Capsule Network

A capsule network is used here to understand the hierarchical relationship between objects and also preserve the spatial relationship. It analyzes the visual data in different conditions of object deformation and rotation. A capsule network aims for the equivariance, not the invariance. Equivariance means changes in the inputs (which are changes in rotation and translation); there will be no predictable changes in the output network while preserving the spatial information. It uses the capsules to store the data and these capsules are the set of neurons. The different layers of the capsule network are the convolutional layer which extracts the features from the input data, the primary layer constructs the capsules and the digit layer predicts the class of the input data based on

the information provided by the capsules. The descriptions of the layers are given below.

Convolutional Layer

The convolutional layer is a building block that is designed to process the data and images. It extracts local features from the input data, such as edges, textures, and patterns, using a collection of learnable filters, or kernels shown in Fig (7). It detects the features from the input and generates the initial input for the capsule network shown in Table (11). Convolution is achieved by applying the convolution operation in a kernel K where the input size is $C_{in} \times H_{in} \times W_{in}$ where C_{in} is the input channels of the input, H_{in} is the height, $and\ W_{in}$ is the width. The shape of the input kernel convolutional is $C_{in} \times K_h \times K_W$ where C_{in} is the depth, K_h is the height, K_W is the width of the kernels and the kernel depth is equal to the input number of channels. The height H_{out} depends on factors like input height H_{in}, and the width W_{out} depends on factor input width W_{in}, The stride of the kernel, and the padding of the input.

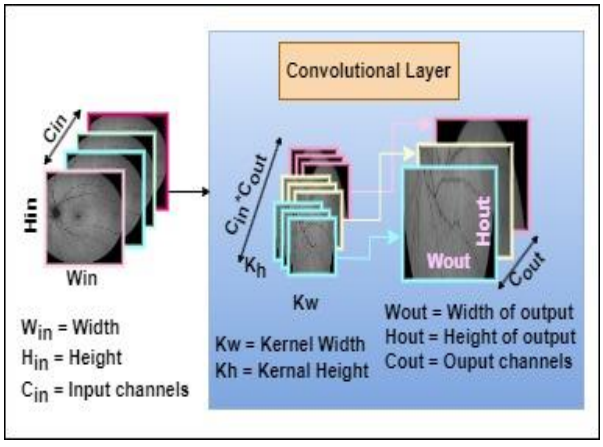


Fig. 7: Local feature extraction

Table 11: Shape of the image after applying the convolutional layer

Stages	Formulas	Output
Reshaping the patch sequence 1D to 2D grid	Total elements = total elements (New)	[1, 4096, 768] = [64, 64, 768]
Applying the convolutional layer	Output height = $(\frac{64-3+2\times1}{1}) + 1 =$ 64	[64, 64, 32]
Calculation of the shape output	Output height = $(\frac{64-3+2\times1}{1}) + 1 =$ 64	[64, 64, 32]

To calculate the height and width Eqs. (9-10) are used:

$$Output\ height\ calculation\ = \left(\frac{Input\ Height - Filter\ Height + 2 \times padding}{Stride}\right) + 1 \quad (9)$$

$$Output\ width\ calculation\ = \left(\frac{Input\ width - Filter\ Width + 2 \times padding}{Stride}\right) + 1 \quad (10)$$

where, input height is the height of the input feature map, filter height is the height of the convolutional filter (kernel), input width is the width of the input feature map, filter width is the width of the convolutional filter (kernel), Stride is the step size of the filter which mover across the input feature map and padding is added to maintain and adjust the output dimensions.

Primary Capsule

The primary capsule converts the raw output of the convolutional layer into the structured representations shown in Fig. (8). Rather than employing scalar activations, the primary capsule uses vector representations to encode low-level features from the input data shown in Table (12).

To calculate the output of the primary capsule, Eq. (11) is used:

$$u_i = ReLU(W_i * X + b_i) \quad (11)$$

where, u_i is the output vector of the i^{th} primary capsule, W_i is the convolutional filter or weight matrix for the i th capsule, * denotes the convolutional network, b_i is the bias term ReLU is the rectified linear unit activation function.

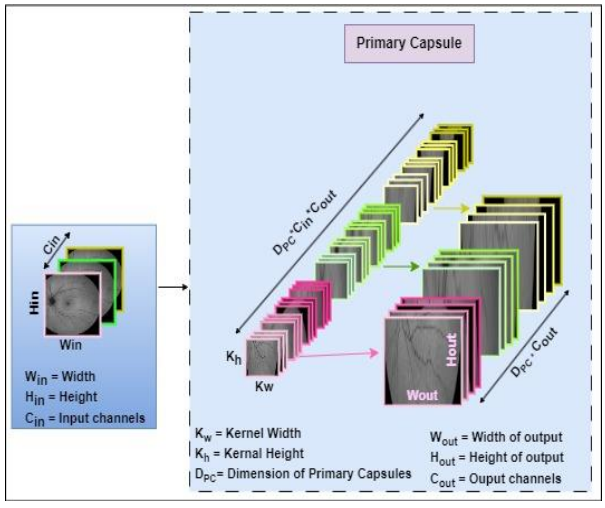


Fig. 8: Primary capsule layers

Table 12: Output of the primary capsule layer

Stages	Formulas	Output
Applying convolutional layers	$u_i = ReLU(W_i * X + b_i)$	4×4×32×8
Output dimension	Output size = $\left(\frac{H-K}{S}\right) + 1$	Where H = 8, K = 9 , S = 2; output height and width is = $\left(\frac{8-9}{2}\right) + 1 = 4$; output shape = 4×4×32×8
Reshaping to capsule form	Reshaped output to size = H'×W'×M × d	4× 4 ×32 × 8
Squash activation	Squash(s) = $\frac{\|\|s\|\|^2}{1+\|\|s\|\|^2}\frac{s}{\|\|s\|\|}$	4× 4 ×32 × 8; squash function scales each vector so that its length is between 0 and 1
Output capsule vector	Each capsule output vector is V_i = squash (u_i)	4× 4 ×32× 8, in the final output primary capsule, consists of 32 capsule vectors of 4× 4 spatial locations

Digit Capsules

Digit capsules recognize and classify the entire digits shown in Fig. (9). It takes the input from the previous layer, the primary capsule layer, and determines which digit is present and which is not. It also handles the pose, scale, and variations in the orientations. In a fully connected network, the output layers are the shape of $N_{class} \times 1$, where N_{class} Shows the number of classes. Every class is represented by a capsule of dimension D_{DC}. The shape of the digit capsule block is $N_{class} \times D_{DC}$. Table (13) shows the calculations and final output of the digit capsules. To calculate the digit capsule Eq. (12) is used:

$$V_j = squash\left(\sum_i c_{ij}\, \hat{u}_{j|i}\right) \quad (12)$$

where, V_j is the output vector, $\hat{u}_{j|i}$ is the vector that predicts the outcome from the i^{th} primary capsule to the j^{th} digit capsule, c_{ij} is the coupling coefficient, and squash (.) is the non-linear squashing function.

To calculate squash Eq. (13) is used:

$$squash(S_i) = \frac{||S_i||^2}{1+||S_i||^2}\frac{S_i}{||S_i||} \quad (13)$$

where, S_i is the squash function of the input vector for the digit capsule, i is the weighted sum of the transformed outputs from the primary capsules, $|S_i||^2$ measures the magnitude or length of the vector, $1 + ||S_i||^2$ adding 1 to the squared length, $\frac{S_i}{||S_i||}$ Maintains the direction of the input vector.

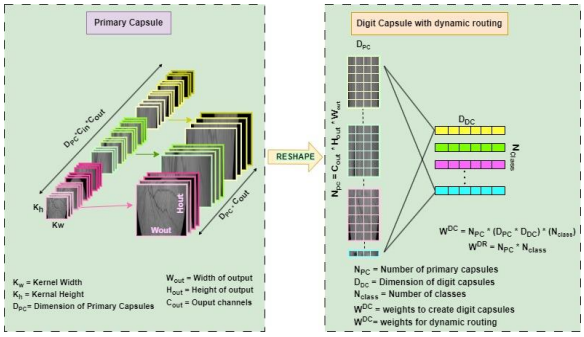


Fig. 9: Digit capsule by dynamic routing

Table 13: The final output of the digit capsule layer

Stages	Formulas	Output
Input	$u_j \in \mathbb{R}^{N\times M\times d_p}$	16× 32 × 8 For every 16 spatial positions, there will be 32 primary capsules
To transfer the output of the primary capsule to the digit capsule initializing the weight	$W_{ij} \in \mathbb{R}^{d_c\times d_p}$, Where *i* belongs to [1, K] and j belongs to [1, M]	10× 32 × 8
Computing the prediction vectors	$\widehat{u_{j\|i}} = W_{ij} \cdot u_j$	[0.3, -0.4, 0.2, 0.1, ⋯ , 0.5]
Initializing the coupling coefficients	$c_{ij} = \frac{1}{k}$	16× 32 × 10
Squash function	squash(S_i) = $\frac{\|\|S_i\|\|^2}{1+\|\|S_i\|\|^2}\frac{S_i}{\|\|S_i\|\|}$	16 × 10 × 16
The final output of the capsule is	$v_i = s_i$	10 × 16

Dynamic Routing

Dynamic routing establishes a more effective connection between the primary capsule (low-level capsule) and the digit capsule (high-level capsule). It creates the strength of connections between each pair of capsules. The digit capsules are extracted from the main capsule using a computation. The weight W^{DC} is trained using backpropagation. N_{pc} is the number of primary capsules, i ∈ $[1, N_{PC}]$ is the index for the primary capsule with dimensions D_{PC} and $j \in [1, N_{class}]$ is the index of the digit capsule with dimensions D_{DC}, w_{ij}^{DC} is shape of $D_{PC} \times D_{DC}$, Individual opinion of i regarding the digit capsule j is shown in Eq. (14).

$$\hat{u}_{j|i} = u_i W_{ij}^{DC} \quad (14)$$

where, u_i is the i^{th} primary capsule, we get an individual digit capsule for each i with a block of shape. $N_{class} \times D_{DC}$, W^{DR} is the weight called the routing weights. These routing weights are updated during the forward pass according to how much each individual digit capsule

agrees with the combined one. The routing weight is W^{DR} is of the shape $N_{PC} \times N_{class}$. The routing weight is first started with zero. The coupling coefficients C_{ij} Shown in Eq. (15):

$$C_{ij} = \frac{exp(W_{ij}^{DR})}{\sum_k exp(W_{ik}^{DR})} \quad (15)$$

where, C_{ij} is the coupling coefficient between the primary capsule and the digit capsule, W_{ij} is the initial log prior probabilities indicate the degree of connection between each primary capsule *j* and each digit capsule *i*.

Classification Layer

This layer transforms the network's learned features into class predictions. A classification model consists of a dense layer producing raw scores for different classes, which are then converted into a probability distribution using an activation function. The model's prediction class is chosen based on the highest probability. The classification layer processes the high-level feature vectors and gives the output as NO DR, Mild DR, Moderate DR, Severe DR, or Proliferative DR shown in Fig. (10):

- No DR: This means that there are no diabetic retinal lesions. These weaker vessels may burst as the illness worsens, resulting in hemorrhages that show up as black spots on the retina.
- Mild DR: It is the earlier stage of DR in this stage the symptoms are often difficult to identify and also difficult to detect without an eye examination
- Moderate DR: These patients present with one to three retinal quadrant hemorrhages, together with cotton wool patches, hard exudates, or venous beading. Within a year, there is a 12-27% chance that they will get PDR.
- Severe DR: Intraretinal Microvascular Abnormalities (IRMA) in one or more quadrants, intraretinal hemorrhages (>20 in each quadrant), or venous beading in two or more quadrants are present in these individuals. The absence of neovascularization, which would suggest PDR, is required for these observations. Macular OCT and fluorescein angiography should be used to follow patients with severe NPDR to identify any Diabetic Macular Edema (DME) or early neovascularization.
- Proliferative DR: These patients have either vitreous, retinal hemorrhages, or neovascularization, but no proliferative diabetic retinopathy that has developed into proliferative DR. For further testing and care, these individuals should be referred immediately to a retina specialist. Laser pan-retinal photocoagulation is typically used to treat peripheral neovascularization. These patients must see a retina expert once a month until their condition stabilizes. They might then be seen every six to twelve months

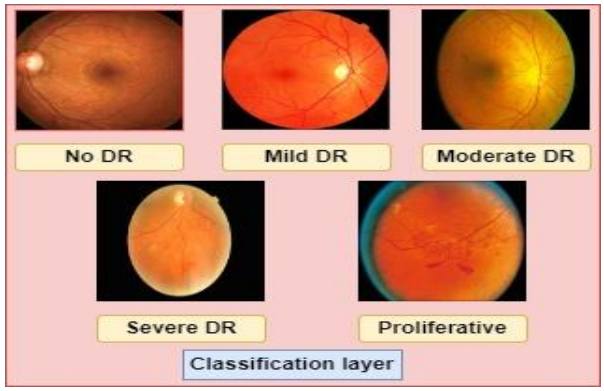


Fig. 10: Classification stages

Results

The result section presents the proposed hybrid ViT-CaspsNet model for DR detection. In this model, the eyepieces dataset is used to train and evaluate the model in which the performance matrices are measured through accuracy, precision, recall, and F1-score. The ViT-CapsNet model gets an accuracy of 94% by comparing the other models which are CNN, ResNet, and EfficientNet. The class-wise performance shows the results for No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR. Visual analysis and graph representation for DR severity and predicted probability, AUC-ROC curves, model performance comparison (for accuracy, precision, and F1-Score), training and validation for Accuracy or loss and the confusion matrix shows the performance of the model more accurately in classifying the images.

Performance Matrix Equations

To diagnose the different classes of the ViT-CapsNet model, the different equations of the performance matrices, accuracy, precision, recall, and F1-score, were used. In which the TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative values, respectively. TP shows the number of cases predicted correctly having the specific DR, TN shows the number of cases correctly predicted but not having the specific DR, FP shows the number of cases incorrectly predicted on a specific DR and FN shows the number of cases incorrectly predicted but not having the specific DR:

- Accuracy
  Accuracy measures the percentage of the correctly identified classes that indicate the model accuracy for both DR stages and health cases. To calculate the accuracy Eq. (16) is used:

  $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

  The model achieved a 94% accuracy rate for classifying the DR images demonstrating its effectiveness and performance

- Precision
  Measuring the model's accuracy at a given DR is called precision; it is expressed as the percentage of true positive predictions among the model's positive predictions. It mainly focuses on the correct positive values. To calculate the precision Eq. (17) is used:

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

  The model achieved a 0.92 precision rate for classifying the DR images demonstrating its effectiveness and performance. This is crucial in high false positive cases due to its reliability in positive predictions
- Recall
  Recall evaluates the accuracy of the model by identifying the positive instances and detecting the relevant cases of the DR classification. To calculate the recall Eq. (18) is used:

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

  The model accurately identifies 91% of true DR classes, demonstrating its effectiveness in detecting DR and minimizing missed cases with a recall of 0.91
- F1-Score
  It combines precision and recall, which measures the model's performance and is particularly used for imbalanced datasets, it also balances the trade-off between these metrics. To calculate the F1-Score Eq. (19) is used:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

  The model achieves an F1-Score of 0.91 which offers a thorough perspective of its overall classification performance

Class Wise Metrics

The model identifies a precision of 0.95 and a recall of 0.93 for No DR. This leads to an imposing F1-score of 0.94 with an accuracy of 95%, demonstrating accuracy. Strong performance in recognizing early-stage disease is demonstrated by the detection of mild DR, for which the precision is 0.88 and recall is 0.85. This results in an F1-Score of 0.86 and an accuracy of 92%. The model reports a recall of 0.80, an F1-score of 0.82, and an accuracy of 89% for moderate DR. Within the Severe DR category, the precision and recall are 0.77 and 0.75, respectively, yielding an F1-score of 0.76 and an accuracy of 84. Finally, the model's precision of 0.70, recall of 0.68, F1-score of 0.69, and accuracy of 80% for Proliferative DR demonstrate the challenges in precisely identifying the disease's most advanced stage, shown in Table (14).

Table 14: Class-wise performance matrices

Classes	Precision	Recall	F1-Score	Accuracy (%)
No DR	0.95	0.93	0.94	95
Mild DR	0.88	0.85	0.86	92
Moderate DR	0.83	0.80	0.82	89
Severe DR	0.77	0.75	0.76	84
Proliferative DR	0.70	0.68	0.69	80

Table 15: Evaluation of each model's performance

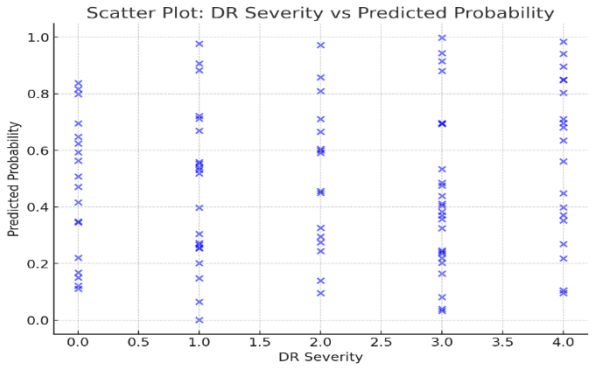
Models	Accuracy (%)	Precision	Recall	F1-Score
CNN	88	0.85	0.82	0.83
ResNet	90	0.88	0.86	0.87
EfficientNet	92	0.90	0.88	0.89
ViT-CapsNet	94	0.92	0.91	0.91

Table (15) illustrates that the ViT-CapsNet model, with an accuracy of 94%, then the other models compared with CNN (88%), ResNet (90%) and EfficientNet (92%). The ViT-CapsNet model appears to correctly classify a greater proportion of images across all the classes, based on its high accuracy. With the lowest percentage, the model shows fewer false positive values having a precision of 0.92. Which will reduce the incorrectly classified images that do not accurately find the true DR cases. The ViT-CapsNet model identifies more accurate cases than the other model by having a recall and F1-score of value 0.91 and 0.91 respectively which shows the model's robustness and demonstrates the balance performance.

Visual Analysis and Graph Representations

The model can be analyzed for performance using different visual representations such as scatter plots, AUC-ROC curves, and comparative analyses. The analysis shows training and validation for accuracy and loss, overfitting in the training. A confusion matrix evaluates the classification outcomes for each DR class. Such visuals are insightful in understanding how robust a model is.

The model's performance in differentiating between DR classes is evaluated using a scatter plot of DR severity versus predicted probability shown in Graph (1). Each point on the graph denotes an individual prediction made by the ViT-CapsNet model. Whereas the x-axis displays the actual severity of the DR class, the y-axis displays the anticipated frequency of the related class. Overestimated probabilities are shown by points above the diagonal and underestimated probabilities are indicated by points below the diagonal.



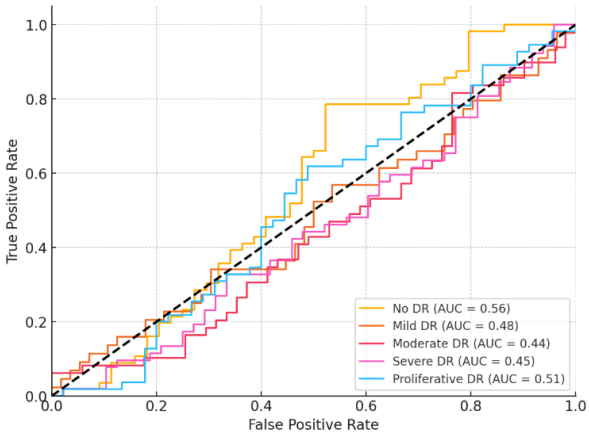
Graph 1: DR severity and predicted probability

Graph (2) shows AUC-ROC scores throughout the evaluation of the model across the various severity categories. Curves showing the true positive rate on the y-axis and the false positive rate on the x-axis are shown for each class. The curve indicates the higher performance of the model when it is close to the top left corner. The No DR suggests the high accuracy of the model which has the AUC likely close to 0.95. AUC values are somewhat lower for Mild DR and Moderate DR, suggesting some compromises between true positives and false positives when identifying intermediate classes of DR. The Severe DR and Proliferative DR classes have the flattest ROC curves, which suggests that the model has a harder time correctly identifying advanced DR classes.

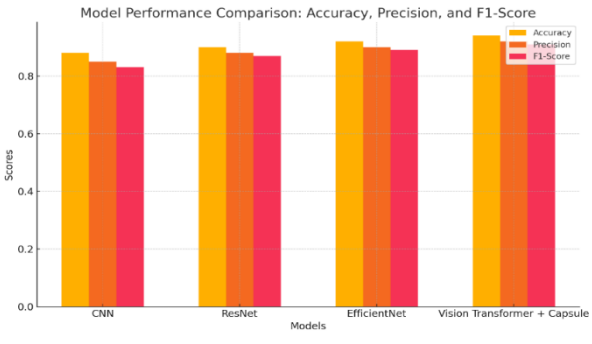
The ViT-CapsNet model gets an accuracy of 94% to the other models this can be done due to the combination of the ViT and the capsule network as shown in Graph (3). The model exceeds in precision achieving with 0.92 score which is higher than the EfficientNet's, ResNet's, and CNNs with scores of 0.90, 0.88, and 0.85 respectively, this increased precision shows that the model can reduce false positive rate. The ViT-CapsNet enhanced precision due to its ability to capture fine-grained features in retinal images. The F1-score of the model which balances the recall and precision is more than the other models that handle the challenging classes of DR or also capture the relevant cases, minimizing missed diagnoses.

The performance of the model can be evaluated through the training and validation accuracy as well as its training and validation scores shown in Graph (4). Validation accuracy measures a model's performance on unseen data, while training accuracy evaluates its ability to accurately classify images within the training dataset. The model reaches 94% of training accuracy which indicates effective learning of retinal features. Validation accuracy indicates the model's generalization which follows the upward trend. The model can demonstrate a strong generalization across various classes of DR, including difficult cases such as Proliferative and Severe DR, as it stabilizes around 92-93%. A well-regularized model that minimizes overfitting and handles unknown data, as well as training data, is suggested by the small difference in accuracy between training and validation runs. Training loss shows how the predictions aligned with the actual labels in the training set. By the end of training, the training loss approaches a low value, suggesting that the model's parameters have been optimized. Although it likewise tends to decrease, the validation loss frequently finds stability at a little greater value than the training loss. High performance is attained, and overfitting and underfitting problems are eliminated when training and validation loss and accuracy curves are aligned.

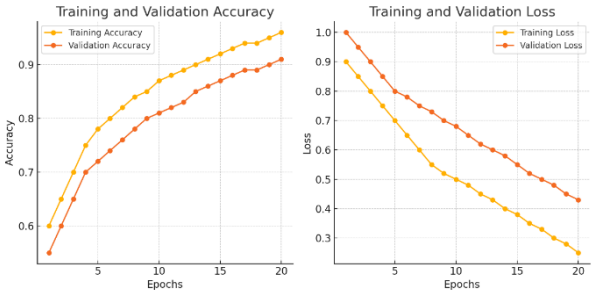
The confusion matrix provides a detailed breakdown of the models in five classes which are No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR shown in Graph (5). In No DR the matrices show the 22116 true positive values which indicates healthy images of correct identifications. In Mild DR, Moderate DR, Sereve DR, and Proliferative DR the matrices have 2106, 4368, 845, and 827 true positives respectively. That indicates model effectiveness, detecting the severe conditions and identifying different classes. The confusion matrix's diagonal elements accurately classify No DR, Mild DR, and Moderate DR, while the off-diagonal elements indicate areas where misclassification occurred.



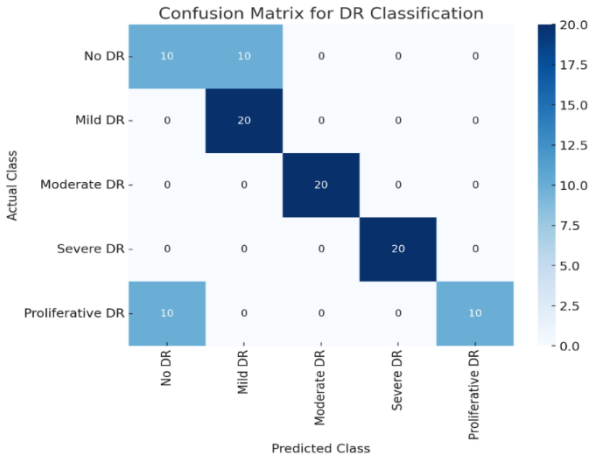
Graph 2: AUC-ROC scores for DR classification by class



Graph 3: Model performance comparison for accuracy, precision, and F1-Score



Graph 4: Training and validation for Accuracy or loss



Graph 5: Confusion matrix for DR classification

Discussion

The ViT-CapsNet model is a novel approach to DR detection, combining Vision Transformers and Capsule Networks. This model achieves an impressive 94% accuracy across five DR classes, reducing false positives, particularly in advanced stages like Severe and Proliferative DR. Its focus on early-stage detection is key, with high F1-scores for No DR and Mild DR, crucial for real-world applications. The model maintains balanced precision and recall values across different DR classes, but its slightly lower performance in Proliferative DR suggests the need for improved optimization techniques. The Vision Transformer's multi-head attention mechanism captures global dependencies, while the Capsule Network ensures robustness against image rotation and deformation. This combination outperforms traditional methods like CNNs and ResNet, which struggle with spatial invariance and lack hierarchical spatial information preservation. However, the model faces challenges in computational demands and reliance on high-quality annotated datasets, which could impact its practical deployment in clinical settings.

Table (16) shows that the ViT-CapsNet model, which is the integration of Vision Transformers and CapsNet models, has achieved 94% accuracy, 0.91 recall, 0.94 F1-Score, and 0.92 precision in DR detection across five classes. It outperforms traditional methods like CNNs and ResNet by capturing global dependencies and preserving spatial relationships. However, challenges include large dataset dependency, high computational demands, and the need for improved interpretability. The model's effectiveness is particularly evident in early-stage detection, with high F1-Scores for No DR (0.94) and Mild DR (0.82). The MAP Concordance Regressive Camargo's Index-Based Deep Multilayer Perceptive Learning Classification (MAPCRCI-DMPLC) model, using the Diabetic Retinopathy Arranged dataset, achieves 92.28% accuracy, focusing on DR classification using advanced neural networks. The IDRIS dataset, using Deep Q Networks (DQN) and Exponential Gannet Pelican Optimization Algorithm (EGFOA), reports a recall of 92.2% but lacks details on accuracy and precision. The EyePACS dataset, using ViT, achieves 88.01% accuracy with high recall and precision. Messidor-2 and APTOS 2019 datasets show accuracies ranging from 87.35-89.35%, focusing on early-stage DR detection and validation. CNN-based model datasets used like Diabetic Retinopathy Detection report lower performance, with accuracy as low as 70% and recall at 50%. The DDR and APTOS dataset CNN512 and YOLOv3 models achieved 89% accuracy with 89% recall. The ViT-CapsNet model stands out for its hybrid approach, ViT for global feature extraction, and CapsNet for spatial hierarchy preservation, providing superior performance across all DR stages, especially in early-stage detection.

Table 16: Comparison with existing techniques

References	Datasets	Techniques	Accuracy in %	Recall in %	F1-Score in %	Precision in %	Classes	Limitations
Muthusamy and Palani (2024)	Diabetic Retinopathy Arranged	MAPCRCI-DMPLC	92.28	-	-	-	Normal, mild, moderate, severe, and proliferative	Enhance DR classification by adopting advanced neural networks for a faster, more accurate, and fully automated grading system to aid screening
Prabhakar *et al*. (2024)	IDRIS	DQN, EGFOA	91.6	92.2	-	-	-	-
Zhang and Chen (2025)	EyePACS Messidor-2 APTOS2019 APTOS2019	ViT	88.01 87.35 89.35 85.34	89.49 89.34 91.71 88.26	89.23 90.10 89.93 86.43	90.12 88.26 89.97 97.34	Normal, Mild, Moderate, Severe,	The research will concentrate on validating the model in various

							and Proliferative	clinical settings to ensure its reliability and explore potential enhancements
Reguant *et al*. (2021)	EyePACS and DIARETDB1	CNN	89~95	74 ~ 86	-	-	No DR, mild NPDR, moderate NPDR, severe NPDR and proliferative DR	The study's performance and assessment may be limited due to its reliance on limited datasets, which lack specific information on DME and laser photocoagulation scars
Khan *et al*. (2021)	EyePacs	Visual Geometry Group-Network in Network (VGG-NiN)	85	55.6	59.6	67	Normal, Mild, Moderate, Severe, PDR	Make significant changes to the existing model's architecture and preprocessing techniques, focusing on the impact on the classification of DR stages, particularly early ones
Sallam *et al*. (2020)	Diabetic Retinopathy Detection	CNN	70	50	-	-	No DR, Mild, Moderate, Severe, Proliferative DR	-
Alyoubi *et al*. (2021)	DDR and APTOS	CNN512 and YOLOv3	89	89	-	-	no-DR, mild, moderate, severe and proliferative DR	Integrate several datasets to attain the dataset's equilibrium
Proposed model	Eyepacs	ViT-CapsNet	94	91	94	92	no DR, mild DR, moderate DR, severe DR, and proliferative DR	The ViT-CapsNet model faces challenges like large dataset dependency, imaging variations sensitivity, high computational demands, inference time, lack of interpretability, and extensive hyperparameter tuning

Conclusion

With an estimated 950,000 affected, DR is the primary cause of vision impairment and blindness in the WHO European Region. This causes damage to the retina's blood vessels, which can result in visual problems and, in certain situations, blindness. If diabetes retinopathy is detected early, it can be prevented. Traditional methods are there to detect the disease but have a low accuracy rate. Hence, the hybrid ViT-CapsNet model is proposed to overcome the challenges and limitations of the traditional model. The model identifies and detects the problem at an early stage. In this model, the eyepieces dataset of 30262 high-resolution fundus images is used

which is the public dataset obtained from Kaggle. Then the data is pre-processed in which the resizing and augmentation process is done to increase the diversity, the data is then split with 70% of the data being a training dataset, 15% being a validation dataset and 15% being a test dataset. ViT extracts the global features and captures the contextual information about the images, while the capsule network preserves the hierarchical connections, both of which are necessary for precise DR classification. This model overcomes the traditional models CNN, ResNet, and EfficientNet. The model is classified into five DR classes that are no DR, mild DR, moderate DR, severe DR, and proliferative DR using performance matrices accuracy, precision, recall, and F1-Score. The model ViT-CapsNet achieves a performance matrices accuracy of 94% and F1-scores of 0.94 for No DR, 0.86 for Mild DR, 0.82 for Moderate DR, 0.76 for Severe DR and 0.69 for Proliferative DR. Comparatively than the other model that is CNN (88% accuracy, F1-score of 0.83), ResNet (90% accuracy, F1-score of 0.87) and EfficientNet (92% accuracy, F1-score of 0.89). Therefore, the ViT-CapsNet helps diabetic patients identify retinal issues early on for a better diagnosis and prevention of vision loss.

Acknowledgment

We are thankful to Chitkara University, Punjab, India for providing support in doing the research work.

Funding Information

No funding is available for doing this research work.

Author's Contributions

Vishal Sharma: Worked on design, methodology and wrote the manuscript.

Rishu: Worked on figures, results and wrote the manuscript.

Vinay Kukreja: Conceptualization, validation and supervision.

Ayush Dogra: Visualization, formal analysis and investigation.

Bhawna Goyal: Visualization and investigation.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

Conflict of Interest

The authors declare that they have no conflict of interest. The authors declare no potential conflicts of interest (financial or non-financial).

References

Al-hazaimeh, O. M., Abu-Ein, A., Tahat, N., Al-Smadi, M., & Al-Nawashi, M. (2022). Combining Artificial Intelligence and Image Processing for Diagnosing Diabetic Retinopathy in Retinal Fundus Images. *International Journal of Online and Biomedical Engineering (IJOE)*, *18*(13), 131–151. https://doi.org/10.3991/ijoe.v18i13.33985

Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning. *Sensors*, *21*(11), 3704. https://doi.org/10.3390/s21113704

Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, *20*, 100377. https://doi.org/10.1016/j.imu.2020.100377

Bajwa, A., Nosheen, N., Talpur, K. I., & Akram, S. (2023). A Prospective Study on Diabetic Retinopathy Detection Based on Modify Convolutional Neural Network Using Fundus Images at Sindh Institute of Ophthalmology & Visual Sciences. *Diagnostics*, *13*(3), 393. https://doi.org/10.3390/diagnostics13030393

Barker, M. M., Davies, M. J., Zaccardi, F., Brady, E. M., Hall, A. P., Henson, J. J., Khunti, K., Lake, A., Redman, E. L., Rowlands, A. V., Speight, J., Yates, T., Sargeant, J. A., & Hadjiconstantinou, M. (2023). Age at Diagnosis of Type 2 Diabetes and Depressive Symptoms, Diabetes-Specific Distress and Self-Compassion. *Diabetes Care*, *46*(3), 579–586. https://doi.org/10.2337/dc22-1237

Bilal, A., Sun, G., Mazhar, S., Imran, A., & Latif, J. (2022). A Transfer Learning and U-Net-based automatic detection of diabetic retinopathy from fundus images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *10*(6), 663–674. https://doi.org/10.1080/21681163.2021.2021111

Bilal, A., Zhu, L., Deng, A., Lu, H., & Wu, N. (2022). AI-Based Automatic Detection and Classification of Diabetic Retinopathy Using U-Net and Deep Learning. *Symmetry*, *14*(7), 1427. https://doi.org/10.3390/sym14071427

Carrera, E. V., Gonzalez, A., & Carrera, R. (2017). Automated detection of diabetic retinopathy using SVM. *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 1–4. https://doi.org/10.1109/intercon.2017.8079692

Das, S., Kharbanda, K., M, S., Raman, R., & D, E. D. (2021). Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. *Biomedical Signal Processing and Control*, *68*, 102600. https://doi.org/10.1016/j.bspc.2021.102600

Dilmurodovna, T. D. (2023). Morphological Signs of the Inflammatory Process in the Pancreas in Type I and Ii Diabetes Mellitus. *European Journal of Innovation in Nonformal Education*, *3*(11), 24–27.

Ebrahimi, B., Le, D., Abtahi, M., Dadzie, A. K., Lim, J. I., Chan, R. V. P., & Yao, X. (2023). Optimizing the OCTA layer fusion option for deep learning classification of diabetic retinopathy. *Biomedical Optics Express*, *14*(9), 4713. https://doi.org/10.1364/boe.495999

ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., Collins, B. S., & Cusi, K. (2023). Introduction and methodology: standards of care in diabetes—2023. *Diabetes Care*, *46*(Supplement_1), S1–S4. https://doi.org/10.2337/dc23-Sint

Franklin, S. W., & Rajan, S. E. (2013). An automated retinal imaging method for the early diagnosis of diabetic retinopathy. *Technology and Health Care*, *21*(6), 557–569. https://doi.org/10.3233/thc-130759

Gao, Z., Pan, X., Shao, J., Jiang, X., Su, Z., Jin, K., & Ye, J. (2023). Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. *British Journal of Ophthalmology*, *107*(12), 1852–1858. https://doi.org/10.1136/bjo-2022-321472

Ghouali, S., Onyema, EM., Guellil, MS., Wajid, M. A., Clare, O., Cherifi, W., & Feham, M. (2022). Artificial Intelligence-Based Teleopthalmology Application for Diagnosis of Diabetics Retinopathy. *IEEE Open Journal of Engineering in Medicine and Biology*, *3*, 124–133. https://doi.org/10.1109/ojemb.2022.3192780

Hameed Abbood, S., Hamed, H. N. A., Mohd Rahim, M. S., M. Alaidi, A. H., & Alrikabi, H. Th. S. (2022). DR-LL Gan: Diabetic Retinopathy Lesions Synthesis using Generative Adversarial Network. *International Journal of Online and Biomedical Engineering (IJOE)*, *18*(03), 151–163. https://doi.org/10.3991/ijoe.v18i03.28005

Islam, M. M., Yang, H.-C., Poly, T. N., Jian, W.-S., & (Jack) Li, Y.-C. (2020). Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, *191*, 105320. https://doi.org/10.1016/j.cmpb.2020.105320

Islam, N., Jony, Md. M. H., Hasan, E., Sutradhar, S., Rahman, A., & Islam, Md. M. (2023). Toward Lightweight Diabetic Retinopathy Classification: A Knowledge Distillation Approach for Resource-Constrained Settings. *Applied Sciences*, *13*(22), 12397. https://doi.org/10.3390/app132212397

Khan, Z., Khan, F. G., Khan, A., Rehman, Z. U., Shah, S., Qummar, S., Ali, F., & Pack, S. (2021). Diabetic Retinopathy Detection Using VGG-NIN a Deep Learning Architecture. *IEEE Access*, *9*, 61408–61416. https://doi.org/10.1109/access.2021.3074422

Liu, J., Gibson, E., Ramchal, S., Shankar, V., Piggott, K., Sychev, Y., Li, A. S., Rao, P. K., Margolis, T. P., Fondahn, E., Bhaskaranand, M., Solanki, K., & Rajagopal, R. (2021). Diabetic Retinopathy Screening with Automated Retinal Image Analysis in a Primary Care Setting Improves Adherence to Ophthalmic Care. *Ophthalmology Retina*, *5*(1), 71–77. https://doi.org/10.1016/j.oret.2020.06.016

Miao, A., Xu, J., Wei, K., Lin, P., Niu, L., Shi, Y., Qian, D., Lu, Y., Jiang, Y., & Zheng, T. (2024). Comparison of B-Scan ultrasonography, ultra-widefield fundus imaging and indirect ophthalmoscopy in detecting retinal breaks in cataractous eyes. *Eye*, *38*(13), 2619–2624. https://doi.org/10.1038/s41433-024-03093-2

Mohan, N. J., Murugan, R., Goel, T., & Roy, P. (2022). ViT-DR: Vision Transformers in Diabetic Retinopathy Grading Using Fundus Images. *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, 167–172. https://doi.org/10.1109/r10-htc54060.2022.9930027

Muthusamy, D., & Palani, P. (2024). Deep learning model using classification for diabetic retinopathy detection: an overview. *Artificial Intelligence Review*, *57*(7), 185. https://doi.org/10.1007/s10462-024-10806-2

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. https://doi.org/10.1109/access.2023.3326528

Nourinia, R., Dastmardi, M., Dastmardi, M., Azimi, R., & Hassanpour, K. (2023). The diagnostic characteristics of direct ophthalmoscopy for diabetic retinopathy screening by family physicians. *International Journal of Diabetes in Developing Countries*, *43*(5), 715–718. https://doi.org/10.1007/s13410-022-01155-3

Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. *Scientific Reports*, *11*(1), 1897. https://doi.org/10.1038/s41598-021-81539-3

Ong, J. X., Konopek, N., Fukuyama, H., & Fawzi, A. A. (2023). Deep Capillary Nonperfusion on OCT Angiography Predicts Complications in Eyes with Referable Nonproliferative Diabetic Retinopathy. *Ophthalmology Retina*, *7*(1), 14–23. https://doi.org/10.1016/j.oret.2022.06.018

Phillip, M., Nimri, R., Bergenstal, R. M., Barnard-Kelly, K., Danne, T., & Hovorka, R. (2023). Consensus recommendations for the use of automated insulin delivery technologies in clinical practice. *Endocrine Reviews*, *44*(2), 254–280. https://doi.org/10.1210/endrev/bnac022

Prabhakar, T., Madhusudhana Rao, T. V., Maram, B., & Chigurukota, D. (2024). Exponential gannet firefly optimization algorithm enabled deep learning for diabetic retinopathy detection. *Biomedical Signal Processing and Control*, *87*, 105376. https://doi.org/10.1016/j.bspc.2023.105376

Priya, R., & Aruna, P. (2013). Diagnosis of diabetic retinopathy using machine learning techniques. *ICTACT Journal on Soft Computing*, *3*(4), 563–575. https://doi.org/10.21917/ijsc.2013.0083

Qomariah, D. U. N., Tjandrasa, H., & Fatichah, C. (2019). Classification of Diabetic Retinopathy and Normal Retinal Images using CNN and SVM. *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, 152–157. https://doi.org/10.1109/icts.2019.8850940

Reguant, R., Brunak, S., & Saha, S. (2021). Understanding inherent image features in CNN-based assessment of diabetic retinopathy. *Scientific Reports*, *11*(1), 9704. https://doi.org/10.1038/s41598-021-89225-0

Roglic, G. (2016). WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), 3–8. https://doi.org/10.4103/2468-8827.184853

Sallam, M. S., Asnawi, A. L., & Olanrewaju, R. F. (2020). Diabetic Retinopathy Grading Using ResNet Convolutional Neural Network. *2020 IEEE Conference on Big Data and Analytics (ICBDA)*, 73–78. https://doi.org/10.1109/icbda50157.2020.9289822

Sebastian, A., Elharrouss, O., Al-Maadeed, S., & Almaadeed, N. (2023). A Survey on Deep-Learning-Based Diabetic Retinopathy Classification. *Diagnostics*, *13*(3), 345. https://doi.org/10.3390/diagnostics13030345

Silberman, N., Ahrlich, K., Fergus, R., & Subramanian, L. (2010). Case for automated detection of diabetic retinopathy. *2010 AAAI Spring Symposium Series*, 85–90.

Skouta, A., Elmoufidi, A., Jai-Andaloussi, S., & Ouchetto, O. (2023). Deep learning for diabetic retinopathy assessments: a literature review. *Multimedia Tools and Applications*, *82*(27), 41701–41766. https://doi.org/10.1007/s11042-023-15110-9

Uppamma, P., & Bhattacharya, S. (2023). A multidomain bio-inspired feature extraction and selection model for diabetic retinopathy severity classification: an ensemble learning approach. *Scientific Reports*, *13*(1), 18572. https://doi.org/10.1038/s41598-023-45886-7

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. https://doi.org/10.3390/s21113883

You, A., Kim, J. K., Ryu, I. H., & Yoo, T. K. (2022). Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. *Eye and Vision*, *9*(1), 6. https://doi.org/10.1186/s40662-022-00277-3

Zaaboub, N., & Douik, A. (2020). Early Diagnosis of Diabetic Retinopathy using Random Forest Algorithm. *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–5. https://doi.org/10.1109/atsip49331.2020.9231795

Zhang, J., & Chen, J. (2024). Research on grading detection methods for diabetic retinopathy based on deep learning. *Pakistan Journal of Medical Sciences*, *41*(1), 225–229. https://doi.org/10.12669/pjms.41.1.9171