# SwinEff-DR: Hybrid Swin Transformer & Efficient Net Architecture for Multi-Scale Diabetic Retinopathy Detection

Anju Mishra[1], Mrinal Pandey[1], Laxman Singh[2]
Manav Rachna University, Faridabad
KIET Group of Institutions, Ghaziabad
E-mail: aanjumishra2108@gmail.com, mrinalpandey@mru.edu.in, laxman.mehlawat01@gmail.com

**Keywords:** diabetic retinopathy, deep learning, convolutional neural networks (CNNs), EfficientNet, swin transformer, medical image classification, SwinEff-DR

**Received:** September 21, 2025

*Diabetic Retinopathy (DR) remains one of the leading causes of preventable blindness worldwide, underscoring the need for early detection and accurate classification. The root cause of this disease is diabetes mellitus. According to the WHO, about 537 million people suffer from the disease and are expected to increase to 783 million by 2047. Diabetic retinopathy (DR) remains an incurable condition; however, early detection can significantly restrict the progression of vision loss. Routine ophthalmic examinations and continuous monitoring play a critical role in preventing blindness associated with DR. Consequently, a pressing need exists for advanced computer-assisted diagnostic systems capable of accurately detecting and grading DR, providing valuable support to ophthalmologists for timely intervention. To address this, we propose SwinEff-DR, a hybrid architecture of Swin transformers with Efficient Network to achieve robust and precise DR classification. The advanced preprocessing on the EyePACS dataset will be performed, then the Swin Transformer as a backbone of the model, followed by Efficient Net. The SwinEff-DR model attains 0.96 precision, 0.97 recall, 0.97 accuracy, and 0.97 F1-score, achieving a 1.49% improvement over existing methods. Furthermore, the framework aligns predictions with standardised severity grading, enabling robust and clinically meaningful diagnostic support.*

*Povzetek: Prispevek predstavlja model SwinEff-DR za zgodnje zaznavanje diabetične retinopatije, ki z visoko natančnostjo izboljšuje klinično diagnostiko.*

## 1 Introduction

Diabetic Retinopathy (DR) is one of the leading causes of preventable blindness worldwide, primarily affecting individuals with long-term diabetes. According to the World Health Organisation, the prevalence of DR continues to rise in parallel with the global increase in diabetes, making its timely and accurate diagnosis a critical healthcare priority [1]. Early detection of DR, particularly in its initial stages marked by microaneurysms and subtle haemorrhages, is vital in preventing vision loss [2]. However, manual grading of fundus images is time-consuming, subjective, and requires trained ophthalmologists, underscoring the need for reliable computer-assisted diagnostic systems [3].

Deep learning, especially Convolutional Neural Networks (CNNs), has demonstrated significant potential in automating DR classification tasks [4], [5]. Despite their effectiveness, CNN-based approaches face several challenges. First, DR lesions occupy less than 1% of fundus pixels, biasing models toward background regions and causing under-detection of tiny yet critical features [6]. Second, class imbalance, particularly the scarcity of severe and proliferative DR cases, further compromises model generalizability [7]. Third, image quality variations, from inconsistent illumination to blur and device-specific differences, hinder accurate lesion identification [8]. Finally, relying on a single CNN architecture often limits robustness, as individual models may overfit specific datasets or fail to generalise across diverse imaging conditions [9].

To address these limitations, a hybrid model has emerged as a promising strategy, leveraging the strengths of two models while reducing individual weaknesses [10].

This study proposes a framework, "SwinEff-DR: Hybrid Swin Transformer & Efficient Net Architecture for Multi-

Scale Diabetic Retinopathy Detection." The framework integrates preprocessing and augmentation to maintain the class imbalance of the dataset. Swin transformer followed by Efficient Net is used to enhance both sensitivity and robustness. The preprocessing techniques (resize) emphasise image fixed-sized management, so all input images must be the same size. Furthermore, data augmentation strategies are employed to address class imbalance and improve generalisation across heterogeneous fundus datasets.

The contributions of this work are summarised as follows:

1. Data Augmentation uses preprocessing pipelines that enhance lesion visibility and provide better feature extraction.
2. A robust swin transformer mechanism that builds hierarchical features, like CNN models and captures both local & global patterns efficiently.
3. Then, an efficient Net architecture is applied as a hybrid-model classifier.
4. The proposed method is evaluated against existing state-of-the-art approaches and demonstrates superior accuracy, precision, recall, and F1-score performance.

By integrating lesion-focused preprocessing with ensemble learning, this framework aims to advance the development of clinically reliable and computationally efficient diagnostic tools for early and accurate DR detection.

## 2 Related work

The studies span a range of methods and datasets in diabetic retinopathy (DR) detection. Ghosh & Chatterjee [11] employed a transfer-learning ensemble of pretrained CNNs (VGG16 and Inception V3), trained on benchmark datasets APTOS, IDRiD, and Messidor-2. Their model achieved 96.4% accuracy, though recall was not evaluated, and lesion-level validation was limited to SCIRParXiv. Bajwa et al. [12] used a modified CNN trained and tested on a private, clinician-labelled fundus-image set collected at the Sindh Institute of Ophthalmology & Visual Sciences (SIOVS)—about 398 patients—and achieved 93.72% accuracy and 97.3% sensitivity. However, model performance varied across datasets and relied heavily on preprocessing MDPIMDPI. For Khan A. et al. [13], details on the specific methodological approach or dataset were not in our search. However, their accuracy was 92%, with unreported precision/recall and a noted issue of dataset imbalance affecting rare DR stages. Hacisoftaoglu et al. [14] used smartphone-acquired images and achieved 98.6% accuracy and 98.2% recall; however, the model's universality remains questionable due to its reliance on smartphone imagery (dataset details unspecified). Khairandish et al. [15] achieved a moderate accuracy of 82% with complete precision and a recall of 81% using a more complex method than a CNN-only model. However, the specifics of the approach and dataset remain unclear. Finally, Jadhav et al. [16] focused solely on ROC-based evaluation with 80.85% accuracy and 76.75% recall, but did not report precision or dataset details, and lacked clinical interpretability. Rajmani et al. [17] used an applied generic deep learning model and optimised hyperparameter tuning using GSCV, achieving 89% accuracy. Zhang. Q. et. al. [18] followed the methodology wherein they combined a preprocessing and segmentation approach before applying Cuckoo search for feature optimisation and obtained commendable results with accuracy metrics surpassing 97.55%.. P.S. Silva et al. [19] proposed a multi-stage deep learning framework for DR grading, integrating lesion segmentation and classification. The model achieved promising results in terms of accuracy. The results obtained on a large DR dataset highlight the potential of deep learning in accurately diagnosing DR. C. Mohanty et al. [20] developed a hybrid model by combining the VGG16 architecture with an XGBoost classifier and further utilising DenseNet-121, achieving an accuracy of 79.50%.

W. L. Alyoubi et al. [21] employed a CNN for DR classification in combination with YOLOv3 for lesion localisation, achieving an overall accuracy of 89%. M.K. Yaqoob et al. [22] employed ResNet-50 for feature extraction in combination with a Random Forest classifier, achieving an accuracy of 75.09%. In contrast, G. Zhang et al. [23] proposed a Multi-Model Domain Adaptation (MMDA) approach with transfer learning, incorporating weighted strategies to enhance performance, pseudo labelling and clustering-based

approaches and achieved 90.6% accuracy on the APTOS dataset.

Table 1 shows the comparative literature review for diabetic retinopathy detection. The key contributions of this work can be summarised as follows. First, a carefully designed preprocessing and data augmentation pipeline is employed to enhance lesion visibility and improve feature extraction. Second, a robust Swin Transformer mechanism is integrated to construct hierarchical representations similar to CNNs while efficiently capturing both local and global retinal patterns. Third, EfficientNet is incorporated within a hybrid classification framework, complementing the transformer features with high-resolution lesion details. Finally, the proposed method is rigorously compared with existing state-of-the-art approaches, demonstrating superior accuracy, precision, recall, and F1-score performance. By combining lesion-focused preprocessing with ensemble-based learning, the framework contributes toward developing clinically reliable and computationally efficient diagnostic tools for early and accurate detection of diabetic retinopathy.

Table 1: Literature review

| **Study / Model (Year)** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **Approach Used** | **Dataset Used** | **Limitations** |
|---|---|---|---|---|---|---|
| Ghosh & Chatterjee (2023) [11] | 96.4 | – | – | Transfer learning with an ensemble of VGG16 + InceptionV3 | APTOS, IDRiD, Messidor-2 | Did not evaluate recall; limited lesion-level validation |
| Bajwa et al. (2023) [12] | 93.72 | – | 97.3 | Customized CNN | Private dataset (SIOVS, Pakistan) | Inconsistent performance across datasets; preprocessing dependency |
| Khan et al. (2024) [13] | 92 | – | – | Deep learning (details unspecified) | Dataset unspecified (imbalanced) | Dataset imbalance; poor generalisation on rare DR stages |
| Hacisoftaoglu et al. (2020) [14] | 98.6 | – | 98.2 | Smartphone-based image classification with CNN | Smartphone-acquired retinal images | Limited to smartphone images; lacks universality |
| Khairandish et al. (2022) [15] | 82.0 | 81.0 | 82.0 | Hybrid model (more complex than CNN-only) | Dataset not specified | Lower accuracy than CNN-only; high training complexity |
| Jadhav et al. (2025) [16] | 80.85 | – | 76.75 | ROC-based analysis (deep learning) | Dataset unspecified | Focuses only on ROC metric; ignores clinical interpretability |

# 3 Proposed methodology

The proposed methodology for diabetic retinopathy detection begins with collecting a fundus image dataset, which is then preprocessed to enhance quality and augmented through techniques like gray scale conversion, rotation, flip and zoom to increase data diversity. Resize brings all the heterogeneous images to the fixed size and maintains the input on the same level. The hybrid model SwinEff-DR extracts the features locally and globally. Fusion of outputs performed, then finally, the system classifies retinal images into five categories, viz., No DR, Mild DR, Moderate DR, Severe NPDR, and PDR, delivering an automated, accurate, and reliable framework for early detection and standardised grading of diabetic retinopathy. Figure 1 shows the flow of methodology.

1 **Fundus image dataset:** The research starts with collecting fundus images from the publicly available dataset Kaggle for diabetic retinopathy detection.
2 **Data preprocessing and augmentation:** The retinal images undergo preprocessing to enhance clarity and maintain consistency before training. This involves resizing images to a standardised resolution, reducing noise, and improving contrast of retinal features for better visibility using CLAHE. Furthermore, data augmentation techniques such as grayscale conversion, image resizing, rotation, and flipping are applied to expand the dataset artificially. Sampling techniques are used to maintain the class imbalance. As images for the NO DR stage are huge in number, the down-sampling technique is needed. Images for the 5$^{th}$ stage PDR are available in very few numbers, so they need up-sampling. Up sampling is performed by augmentation.
3 **Hybrid model architecture:** Feature extraction performed in two stages: (1) Local feature extraction, and (2) Global feature extraction. Efficient Net branch extracts the local features and swin transformer branch extracts the global features from the retinal images. The Efficient Net and Swin Transformers outputs are then concatenated using fully connected layers. Both model branches contribute unique strengths in feature extraction and pattern recognition, and their combined predictions minimise misclassification. This ensemble approach significantly improves robustness, accuracy, and reliability compared to individual models.
4 **Classification:** In the final stage, the system assigns each fundus image to one of five categories reflecting the severity of diabetic retinopathy: No DR, Mild DR, Moderate DR, Severe NPDR, or PDR. SwinEff-DR models' classification supports ophthalmologists by clearly indicating disease progression and guiding timely treatment decisions.

---

**Algorithm:** Hybrid Swin Transformer–EfficientNet (SwinEff-DR)

---

**Input:** Fundus image dataset $D=\{x_i,y_i\}_{i=1}^{N}$, where $x_i$ is an image and $y_i \in \{0,1,2,3,4\}$ represents DR stage.
**Output:** Predicted class probabilities for 5 DR stages

**Step 1:** Data Preprocessing
{
(i) Resize all images to 512×512.
(ii) Apply CLAHE for contrast enhancement.
(iii) Apply augmentation: resize, rotation, and flipping.
}

**Step 2:** EfficientNet Branch (Local Features)
{

(i) Input $x_i$ into the EfficientNet-B4 backbone.
(ii) Apply:
- Stem Convolution + BatchNorm
- MBConv blocks with depthwise separable convolutions
- Squeeze-and-Excitation layers

(iii) Output: Local lesion feature map $F_{eff}(x_i)$

}

**Step 3:** Swin Transformer Branch (Global Features)
{
(i) Partition the image into non-overlapping patches (4×4).
(ii) Linear embedding of patches → input tokens.
(iii) For each stage:
- Apply Window-based Multi-Head Self-Attention (W-MSA).
- Apply Shifted Window Attention (SW-MSA) for cross-window learning.
- Downsample to build a hierarchical representation.

(iv) Output: Global context feature map $F_{swin}(x_i)$.
}

**Step 4:** Feature Fusion
{
(i) Concatenate $F_{eff}(x_i)$ and $F_{swin}(x_i)$.
(ii) Pass through dense layers.
(iii) Apply attention-based fusion for weighted feature importance.
(iv) Apply Batch Normalisation + Dropout for regularisation.
(v) Output: Fused feature vector $F_{fusion}(x_i)$.
}

**Step 5:** Classification Head
{
(i) Feed $F_{fusion}(x_i)$ into fully connected layers.
(ii) Apply **Softmax** activation.
(iii) Output probability distribution across 5 DR stages:

$$P(y/x_i) = Softmax(W. F_{fusion}(x_i)+b)$$

}

**Step 6:** Training
{
(i) Define loss function: Weighted categorical cross-entropy.
(ii) Use optimiser: AdamW with learning rate scheduling.
(iii) Train for $E$ epochs with batch size $B$.
(iv) Apply early stopping if validation loss does not improve.
}
**End**

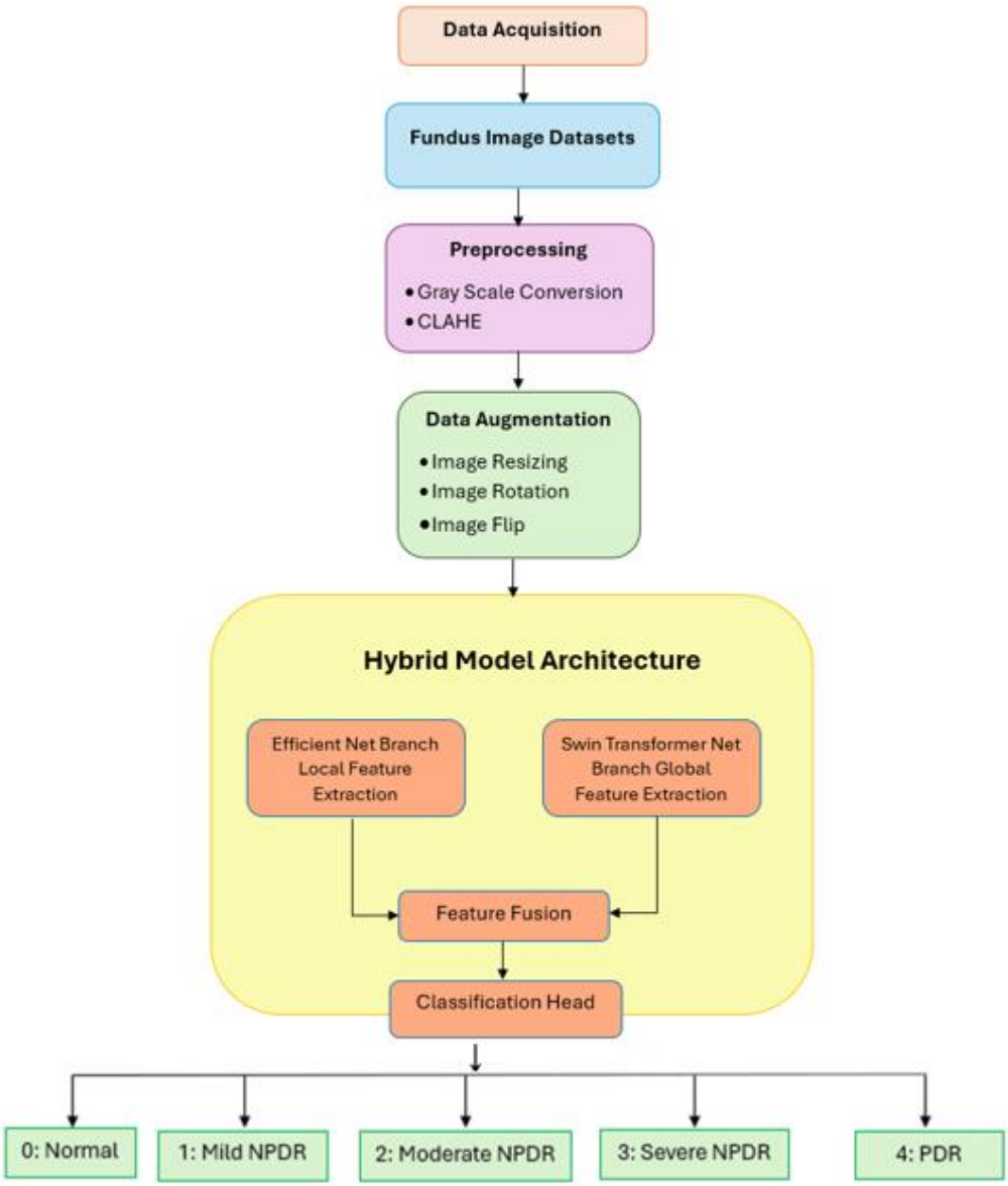


Figure 1: Flow of methodology

## 3.1 Dataset used

Publicly available retinal fundus image datasets have played a vital role in advancing automated diabetic retinopathy (DR) detection. These datasets provide retinal scans from patients at different stages of DR, making them essential for training and evaluating computer-aided diagnostic systems. Such systems utilise advanced imaging.

Processing techniques and machine learning algorithms to ensure accurate disease classification. This study used the EyePACS dataset, an open-source collection available on

Kaggle, comprising 35,126 retinal fundus images. The dataset is categorised into five classes: No DR, Mild DR, Moderate DR, Severe NPDR, and Proliferative DR (PDR). The class-wise distribution of images is summarised in Table 2

Table 2: Details of fundus image dataset

| Class Label | Stage of DR/ Classes | Image Count | Image count after Down/Up Sampling |
|---|---|---|---|
| 0 | No DR | 25376 | 3500 |
| 1 | Mild DR | 2495 | 2495 |
| 2 | Moderate DR | 5476 | 5476 |
| 3 | Severe NPDR | 1013 | 2500 |
| 4 | PDR | 765 | 2200 |

## 3.2 Image preprocessing and augmentation

To enhance the robustness and generalizability of the SwinEff-DR model, a comprehensive preprocessing and augmentation pipeline was applied to the retinal fundus images. Preprocessing involved grayscale conversion and applying contrast enhancement techniques such as CLAHE (Contrast Limited Adaptive Histogram Equilization) to improve the visibility of retinal structures. In addition, data augmentation techniques such as resizing, rotation, flipping, and scaling all images to a uniform resolution, and denoising to remove background artefacts were performed to expand the dataset artificially. These augmentations increased dataset diversity and reduced the risk of overfitting, enabling the model to generalise better across variations in orientation, scale, and image quality.

### 3.2.1 Gray scale conversion

The grayscale conversion process transforms an image from its original color space (such as RGB, CMYK, or HSV) into shades of gray, ranging from black to white [24]. This reduction in colour information simplifies the image while preserving essential structural details, facilitating more efficient feature extraction for subsequent processing steps. These shades range between completely white and black and are calculated using equation (1):

Gray Scale=0.299R+0.587G+0.114B……..(1)

### 3.2.2 Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE is an image enhancement technique that improves local contrast by operating on the image's small regions, or tiles. Unlike standard histogram equalization, CLAHE limits contrast amplification through a clipping threshold, preventing noise over-enhancement. The enhanced tiles are combined smoothly using bilinear interpolation, resulting in a uniform image with improved visibility of subtle features. In medical imaging, such as fundus analysis, CLAHE effectively highlights microaneurysms, hemorrhages, and exudates, facilitating more accurate segmentation and classification.

### 3.2.3 Image resizing

To ensure uniformity of input dimensions across the dataset, all images were resized to a fixed resolution (e.g., 512 × 512 pixels). The resizing transformation can be expressed by equation (2):

$$I'(x', y') = I\left(\frac{x'}{s_x}, \frac{y'}{s_y}\right) \quad \text{……(2)}$$

Where $S_x$, $S_y$ are the scaling factors along the width and height axes [25]. This normalisation allows the images to be directly utilised by a convolutional neural network architecture without distortion.

### 3.2.4 Image rotation

Fundus datasets are often imbalanced and limited. Rotating images generates new training samples without collecting extra data. In real-world screening, fundus images can be captured at slightly different angles. Rotation augmentation helps the model remain invariant

to orientation changes [26]. In image rotation, the input image is rotated by a certain angle (i.e., angles between -30 ° to +30°)

### 3.2.5 Image flip

Image flipping is of two types, i.e., Horizontal flipping and vertical flipping. In these techniques, fundus images are flipped left-right and up-down, which increases symmetry variations and prevents overfitting. Figure 2 represents every image state after applying all the image preprocessing techniques.

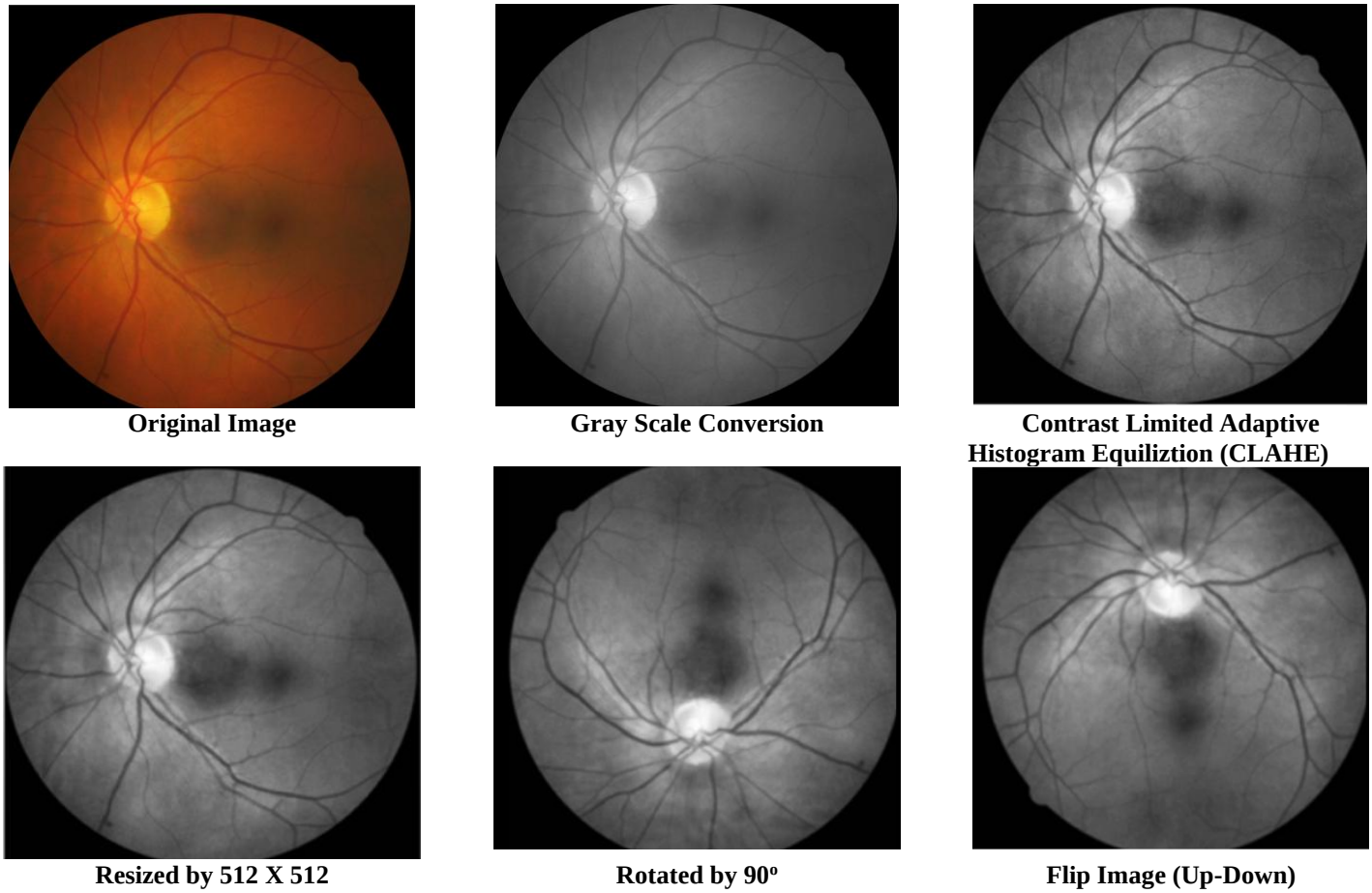


Figure 2: Conversion of fundus image after image preprocessing

## 3.3 Hybrid model architecture

The hybrid Swin Transformer–EfficientNet architecture leverages the complementary strengths of convolutional neural networks (CNNs) and vision transformers, enabling robust and reliable detection of diabetic retinopathy (DR). In the proposed model, EfficientNet functions as a lightweight CNN branch that captures fine-grained local lesion features such as microaneurysms, haemorrhages, and exudates. In contrast, the Swin Transformer branch extracts global contextual information of the retina through hierarchical shifted-window attention. In the proposed model, EfficientNet functions as a lightweight CNN branch that captures fine-grained local lesion features such as microaneurysms, haemorrhages, and exudates. In contrast, the Swin Transformer branch extracts global contextual information of the retina through hierarchical shifted-window attention.

### 3.3.1 Swin transformer

The Swin Transformer is a vision transformer (ViT) model that introduces a hierarchical design and shifted window attention to process high-resolution images efficiently [27]. Unlike traditional Vision Transformers (ViTs), which compute global self-attention across all image patches and incur high computational cost, the Swin Transformer introduces window-based multi-head self-attention (W-MSA), where self-attention is restricted to non-overlapping local windows [28]. To enable cross-window communication and capture global context, it employs shifted window attention (SW-MSA), where the windows are shifted by half their size in alternate layers. The architecture builds a hierarchical representation by gradually merging patches, like CNN downsampling, allowing the model to capture fine-grained local features (e.g., microaneurysms or exudates in fundus images) and global contextual information (disease progression across the retina). This makes the Swin Transformer highly effective for diabetic retinopathy detection, as it balances efficiency with strong feature learning capabilities for multi-scale lesion analysis.

Patch Partitioning and embedding is defined as an input image $I \in R^{H\times W\times 3}$, *which* is divided into non-overlapping patches of size $P\times P$. Each patch is flattened and linearly projected to obtain tokens as in the equation ():

$$X_0=\{x_1, x_2,\ldots,x_N\}, \qquad x_i \in R^C \ldots\ldots\ldots\ldots\ldots()$$

where $N=H\times W/ P^2$ is the number of patches and $C$ is the embedding dimension [29].

Multi-Head Self-Attention (MSA) for each window tokens $X \in R^{M2\times C}$ (where M is the window size), self-attention is computed as:

$$\text{Attention}(Q,K,V)=Softmax(((QK^T)/\sqrt{d_k})+ B)V \ldots.()$$

Where, $Q=XW_Q$, $K=XW_K$, $V=XW_V$

$W_Q, W_K, W_V \in R^{C\times dk}$ are learnable projection matrices, and BBB is the relative positional bias added to capture spatial structure.

Window-based Self-Attention (W-MSA) is restricted to non-overlapping local windows of size $M\times M$, instead of computing attention across the entire image. This reduces computational complexity by giving equation (3):

$$O((HW)^2) \text{ (Global MSA)} \rightarrow O(M^2HW) \text{ (W-MSA)}\ldots..(3)$$

Shifted Window Self-Attention (SW-MSA) enables information flow between adjacent windows; the window partitioning is shifted by $\lfloor M/2 \rfloor$ pixels in alternating layers [30]. Thus, the output feature of a Swin Transformer block is defined as:

$$Z^{l+1} = \text{MLP}(\text{LayerNorm}(\text{SW-MSA}(Z^l)))\ldots\ldots\ldots\ldots(4)$$

$$Z^{l+2}=\text{MLP}(\text{LayerNorm}(\text{W-MSA}(Z^{l+1})))\ldots\ldots\ldots\ldots(5)$$

Where $Z^l$ is the input to the $l$-th layer, and MLP denotes a feed-forward network with residual connections.

In hierarchical representation, at each stage, non-overlapping patch merging reduces the resolution by a factor of 2 while doubling the embedding dimension $C$, producing multi-scale features defined by equation (6):

$$(H, W, C) \rightarrow \left(\frac{H}{2}, \frac{W}{2}, 2C\right)\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Where $H$ and $W$ denote the input height and width, $C$ is the initial embedding dimension [31].

### 3.3.2 Efficient Net Model

EfficientNet represents a family of convolutional neural networks (CNNs) designed to deliver high accuracy while preserving computational efficiency through a compound scaling strategy [32]. In contrast to traditional models that scale along a single dimension, such as depth, width, or input resolution, EfficientNet uniformly scales all three dimensions in a balanced manner, optimising both performance and efficiency. EfficientNet introduces compound scaling, where all three dimensions are scaled in a balanced manner using a set of fixed coefficients baseline network, EfficientNet-B0, is constructed using mobile inverted bottleneck convolution (MBConv) blocks integrated with squeeze-and-excitation (SE) modules, which enhance feature recalibration and improve the representational capacity of the model [33].

Each MBConv block consists of four main steps:

1. **Expansion phase:** The input channels are expanded using a pointwise (1×1) convolution, followed by the Swish activation function.
2. **Depthwise convolution:** A depthwise separable convolution captures spatial features while maintaining efficiency.
3. **Squeeze-and-Excitation (SE):** A global average pooling operation squeezes channel information into a vector, passing through fully connected layers to adaptively reweight feature

channels, enhancing informative features and suppressing less relevant ones.

4. **Projection phase:** The output is projected back to a lower-dimensional representation using another 1×1 convolution, often combined with a residual skip connection for stable training.

By stacking multiple MBConv blocks and progressively reducing spatial resolution while increasing channel depth, EfficientNet constructs a hierarchical feature representation similar to traditional CNNs. The final feature maps are aggregated using global average pooling, followed by fully connected layers for classification [34].

Using the compound scaling rule, EfficientNet achieves consistent performance improvements across multiple model sizes (B0–B7) without compromising efficiency. This property makes it particularly advantageous for medical imaging tasks such as diabetic retinopathy detection, where the simultaneous recognition of subtle lesion-level abnormalities (e.g., microaneurysms, haemorrhages) and broader retinal structures is required, ensuring minimal computational overhead.

### 3.3.3 SwinEff-DR model

The proposed hybrid model integrates the EfficientNet and the Swin Transformer to leverage both local lesion-level features and global contextual representations for accurate multi-stage diabetic retinopathy (DR) classification. On one branch, EfficientNet is a convolutional backbone that extracts fine-grained, local retinal features through stacked MBConv blocks combined with squeeze-and-excitation (SE) modules. This allows the network to highlight salient pathological patterns such as microaneurysms, exudates, and haemorrhages while maintaining computational efficiency through depthwise separable convolutions and compound scaling.

On the parallel branch, the Swin Transformer processes the image as a sequence of non-overlapping patch embeddings. Using window-based multi-head self-attention (W-MSA), the model captures local dependencies efficiently, while shifted window attention. (SW-MSA) enables cross-window interactions for global context learning. Swin constructs multi-scale feature representations through hierarchical patch merging, providing a broader understanding of disease progression across the retina.

The outputs of both branches, local feature maps from EfficientNet and global feature maps from the Swin Transformer, are aligned in dimension and fused via either concatenation with dense projection or an attention-based fusion mechanism that adaptively weights each modality. The fused representation is passed through fully connected layers, followed by a softmax classifier, to output probability distributions over the five DR stages.

This integration combines the strengths of CNNs (efficient local feature extraction) with the advantages of transformers (long-range dependency modelling and scalability), resulting in a well-suited model for multi-scale lesion detection and robust DR grading.

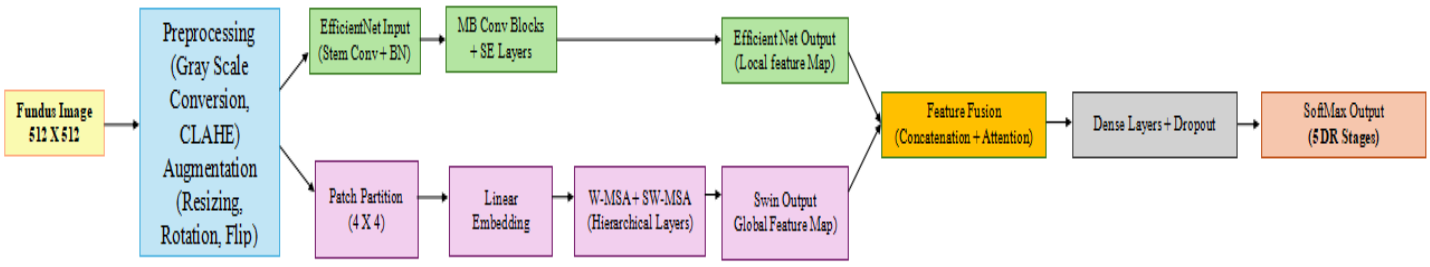


Figure 3: SwinEff-DR

Figure 3 shows the architecture of the SwinEff-DR model. When comparing the EfficientNet Branch with the Swin Transformer Branch, the advantage lies in utilising Fusion + classification. The proposed Hybrid SwinEff-DR model combines EfficientNet and the Swin Transformer, which are jointly employed to capture both local and global retinal features for diabetic retinopathy (DR) classification. Preprocessed fundus images (512×512, CLAHE-enhanced) are fed in parallel into two branches: the EfficientNet branch, which leverages MBConv blocks, depthwise separable convolutions, and squeeze-and-excitation modules to extract fine-grained lesion features, and the Swin Transformer branch, which partitions images into 4×4 patches and applies window-

based as well as shifted window self-attention to model long-range dependencies and hierarchical global structures. The feature representations from both branches are concatenated and refined through an attention-based fusion module, followed by batch normalisation and dropout for regularisation. Finally, a fully connected layer with softmax activation classifies the images into five DR stages: No DR, Mild, Moderate, Severe, and Proliferative.

Table 3: Stage description of models

| **Stage** | **EfficientNet Branch (Local Features)** | **Swin Transformer Branch (Global Features)** | **Fusion + Classification** |
|---|---|---|---|
| **Input** | Fundus Image (512×512) → Gray Scale Conversion, CLAHE, augmentation | Fundus Image (512×512) → Gray Scale Conversion, CLAHE, augmentation | – |
| **Stage 1** | Stem Conv + BatchNorm | Patch Partition (4×4) + Linear Embedding | – |
| **Stage 2** | MBConv Blocks (MBConv1, MBConv6) | Window-based Multi-head Self Attention (W-MSA) + MLP | – |
| **Stage 3** | MBConv Blocks with Depthwise Separable Convs + Squeeze-and-Excitation | Shifted Window Attention (SW-MSA) + Downsampling | – |
| **Stage 4** | Deeper MBConv blocks (wider + deeper layers for fine lesion features) | Hierarchical Attention (Stage 3 & Stage 4 feature maps) | – |
| **Feature Extraction** | High-resolution local lesion features (microaneurysms, haemorrhages) | Global multi-scale retinal patterns (vessel distortion, lesions) | Concatenation of EfficientNet + Swin outputs |
| **Fusion Module** | – | – | Dense Layers + Attention-based fusion |
| **Regularization** | – | – | BatchNorm + Dropout |
| **Classification Head** | – | – | Fully Connected Layers + Softmax → 5 DR Stages (No, Mild, Moderate, Severe, Proliferative) |

Table 3 represents the Stage-wise architecture of the proposed SwinEff-DR model. The EfficientNet branch extracts local lesion features (e.g., microaneurysms, haemorrhages) through MBConv and SE layers. In contrast, using hierarchical attention, the Swin Transformer branch captures global contextual patterns (e.g., vessel distortions, multi-scale lesions).

These complementary features are fused via concatenation and attention, followed by dense layers and a SoftMax classifier to predict the five DR stages, i.e., No DR, Mild DR, Moderate DR, Severe NPDR, and Proliferative DR.

Table 4: Tuning of Hyparparameter

| **Hyperparameter** | **Search Range** | **Optimal Value (Tuned)** | **Remarks** |
|---|---|---|---|
| **Learning Rate (LR)** | 1e-5 → 1e-3 | **3e-4** | Cosine annealing scheduler used |
| **Batch Size** | 16, 32, 64 | **32** | Balanced efficiency vs. GPU memory |
| **Optimizer** | Adam, AdamW, SGD | **AdamW** | Stable convergence, better generalisation |
| **Weight Decay** | 0.001 → 0.1 | **0.01** | Prevents overfitting |

| | | | |
|---|---|---|---|
| **Dropout (EfficientNet)** | 0.1 → 0.5 | **0.3** | Reduces overfitting in dense layers |
| **Stochastic Depth (Swin)** | 0.05 → 0.3 | **0.2** | Improves robustness |
| **Patch Size (Swin)** | 4, 8 | **4** | Preserves lesion-level features |
| **Window Size (Swin W-MSA)** | 4, 7, 8 | **7** | Best balance of local vs. global context |
| **Fusion Strategy** | Concatenation, Weighted Sum, Attention-based | **Attention-based Fusion** | Learns adaptive weighting of features |
| **Loss Function** | Cross-Entropy, Weighted CE, Focal Loss | **Focal Loss (γ=2, α-balanced)** | Handles class imbalance |
| **Epochs** | 50 → 150 | **100** (with Early Stopping, patience=12) | Prevents overfitting |
| **Data Augmentation** | Rotation (±5°–20°), Resize and Flip, | **Rotation ±15°, Flip,(±20%)** | Preserves medical realism |

The hyperparameter tuning process for the proposed SwinEff-DR model is summarised in Table 4. The optimal learning rate was set to *3e-4* with a cosine annealing scheduler, enabling smooth convergence during training. A batch size of 32 was chosen as the best compromise between computational efficiency and GPU memory usage. Among the optimisers tested, AdamW proved most effective, offering stable convergence and improved generalisation compared to Adam and SGD. Regularisation played a key role, with a weight decay of 0.01 and dropout of 0.3 in EfficientNet layers helping to mitigate overfitting. At the same time, a stochastic depth rate of 0.2 in the Swin branch improved model robustness. For the Swin Transformer, a patch size of 4 preserved fine lesion-level details, and a window size of 7 provided an optimal balance between local and global context. An attention-based fusion strategy was adopted to weight local and global features adaptively, outperforming simple concatenation or weighted summation. Focal loss (γ=2, α-balanced) was selected over cross-entropy variants to address class imbalance. Training was conducted for up to 100 epochs with early stopping (patience=12) to prevent overfitting. Data augmentation, including ±15° rotation, resizing, and ±20% flipping, was applied to increase dataset variability while maintaining medical realism. These tuned hyperparameters ensured robust learning, reduced overfitting, and enhanced model generalisation across the five DR stages.

# 4 Evaluated results

To assess the computational overhead of the proposed **SwinEff-DR** model, which integrates EfficientNet and Swin Transformer to extract complementary retinal features. The comparative performance of different models for diabetic retinopathy classification is summarised in Table 5. The results indicate that the Swin Transformer achieved an accuracy of 94.3%, with precision, recall, and F1-score values of 0.953, 0.941, and 0.9432, respectively. The EfficientNet model performed slightly better, reaching an accuracy of 95.12% and balanced precision, recall, and F1-scores of approximately 0.961–0.962. The highest performance was observed with the proposed SwinEff-DR (Fusion + Classification) model, which

Table 5: Performance of individual classification models and SwinEff-DR model

| S. No. | Model | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| 1 | Swin Transformer | 0.953 | 0.941 | 0.9432 | 0.943 |
| 2 | Efficient Net | 0.9612 | 0.9621 | 0.9612 | 0.9512 |
| **3** | **SwinEff-DR (Fusion + Classification)** | **0.9626** | **0.9738** | **0.9732** | **0.9721** |

Integrates local feature extraction from EfficientNet with global context modelling from the Swin Transformer. This hybrid architecture achieved the highest overall accuracy of 97.21%, along with superior precision (0.9626), recall (0.9738), and F1-score (0.9732).

These results demonstrate the effectiveness of combining local and global feature representations, enabling more accurate recognition of subtle retinal abnormalities across different DR stages.

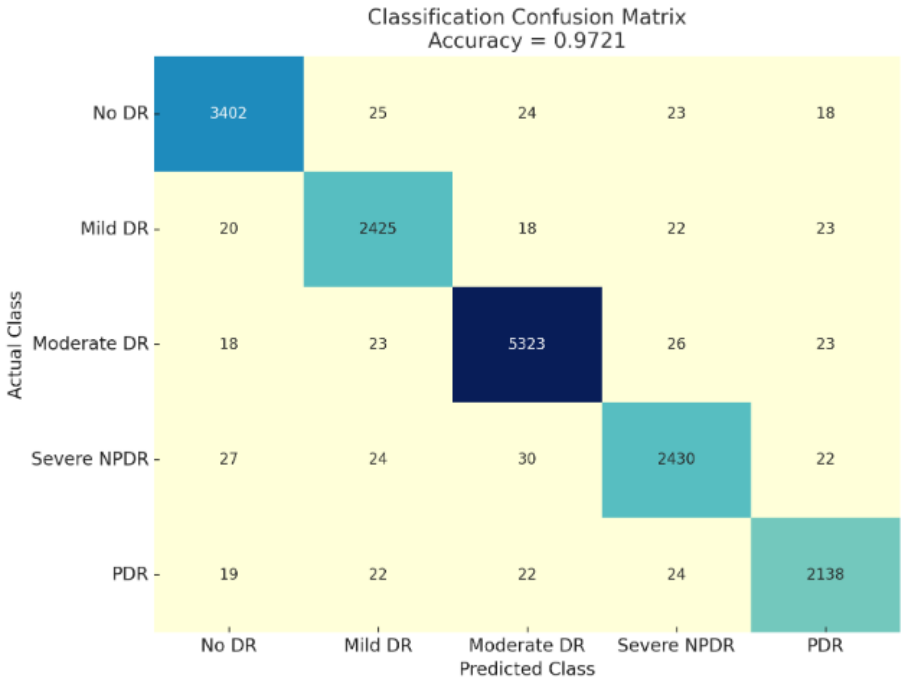


Figure 4: Confusion Matrix of SwinEff-DR

The confusion matrix presented in Figure 4 demonstrates the classification performance of the proposed SwinEff-DR model across five diabetic retinopathy (DR) categories. The model achieved an overall accuracy of 97.21%, with consistently high true positive counts across all classes. Notably, No DR (3402), Mild DR (2425), Moderate DR (5323), Severe NPDR (2430), and PDR (2138) were correctly classified with minimal misclassifications. Most misclassifications occur between adjacent severity levels, such as Mild DR being confused with Moderate DR or Severe NPDR with PDR, which can be attributed to the subtle clinical similarities between consecutive stages. Notably, extreme misclassifications, for instance, No DR being classified as PDR or vice versa, were rare, highlighting the robustness of the model in differentiating healthy cases from advanced pathological stages.

These results validate the model's effectiveness in handling multi-class DR classification and highlight its potential applicability in real-world clinical screening scenarios.

Table 6: Class-wise evaluated results

| Class | Precision | Recall | F1-Score | Support | Accuracy |
|---|---|---|---|---|---|
| **No DR** | 0.9759 | 0.9742 | 0.9751 | 3492 | 0.9731 |
| **Mild DR** | 0.9627 | 0.9669 | 0.9648 | 2508 | 0.9712 |
| **Moderate DR** | 0.9826 | 0.9834 | 0.9830 | 5413 | 0.9961 |

| | | | | | |
|---|---|---|---|---|---|
| **Severe NPDR** | 0.9624 | 0.9593 | 0.9609 | 2533 | 0.9603 |
| **PDR** | 0.9613 | 0.9609 | 0.9611 | 2225 | 0.9601 |
| **SwinEff-DR (AVG)** | **0.9689** | **0.9709** | **0.9689** | **16,171** | **0.9721** |

The classification report in Table 6 and the confusion matrix in Figure 4 collectively validate the robustness of the proposed SwinEff-DR model. The table highlights consistently high precision, recall, and F1-scores across all five classes, with Moderate DR achieving the highest F1-score of 98.30% and accuracy of 99.63%, reflecting the model's strong ability to identify mid-stage disease patterns. Similarly, the confusion matrix reinforces these findings by showing large diagonal values, indicating correct classifications, while the relatively few off-diagonal entries represent minimal misclassifications. Most errors occurred between adjacent stages, such as Mild vs. Moderate DR or Severe NPDR vs. PDR, which is consistent with the clinical overlap of features in these categories. Notably, the confusion matrix and the classification table demonstrate that extreme misclassifications (e.g., No DR labelled as PDR) are rare, underscoring the model's reliability in differentiating healthy from advanced pathological cases. Together, these results confirm that the SwinEff-DR model achieves a global accuracy of 97.21% and maintains balanced performance across all classes, ensuring clinical applicability for early screening and accurate staging of diabetic retinopathy. A comparative analysis of several methods or models based on different performance metrics is presented in Table 7. SwinEff-DR outperforms the other existing models in precision and accuracy, as shown in Figure 5. This proposed model, SwinEff-DR, enhances the accuracy and achieves an improvement of 1.49% from the existing models

Table 7: Performance comparison of SwinEff-DR with existing models

| References | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| [19] | 0.93 | 0.92 | 0.92 | 0.92 |
| [20] | 0.67 | 0.56 | 0.61 | 0.83 |
| [21] | 0.98 | 0.95 | 0.97 | 0.96 |
| [22] | 0.95 | 0.93 | 0.94 | 0.95 |
| [23] | 0.96 | 0.96 | 0.97 | 0.97 |
| [24] | 0.96 | 0.96 | 0.96 | 0.96 |
| [25] | 0.97 | 0.97 | 0.97 | 0.97 |
| **SwinEff-DR (Proposed Model)** | **0.9626** | **0.9738** | **0.9732** | **0.9721** |

The comparison shows that while earlier methods such as [19] and [20] achieved moderate performance, recent approaches ([21]–[25]) reported higher accuracies between 0.96 and 0.97. The proposed **SwinEff-DR model** demonstrates competitive results, with a precision of 0.9626, recall of 0.9738, F1-score of 0.9732, and accuracy of 0.9721, confirming its effectiveness in diabetic retinopathy classification.

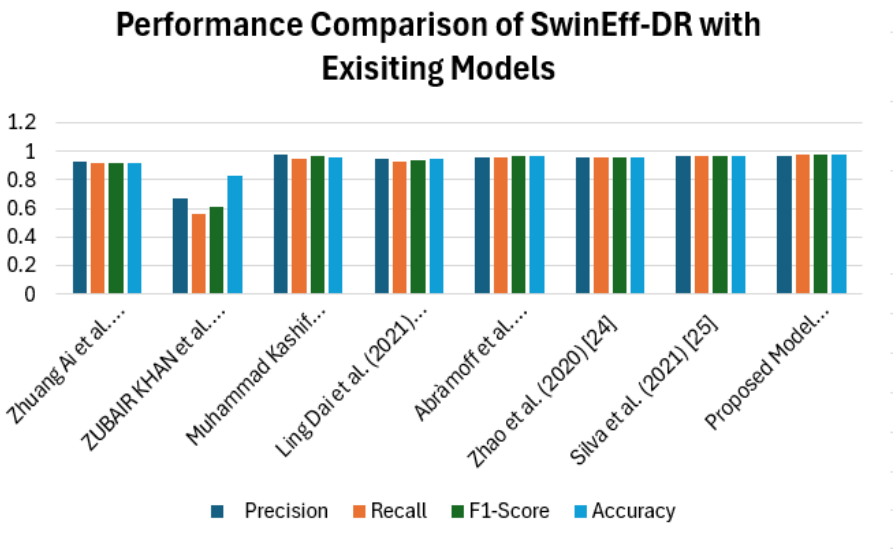


Figure 5: Performance comparison of SwinEff-DR with existing models

# 4 Discussion

Experimental results show that the proposed SwinEff-DR model achieves an overall accuracy of 97.21% for diabetic retinopathy (DR) classification, demonstrating strong reliability in distinguishing among the five stages: No DR, Mild DR, Moderate DR, Severe NPDR, and PDR. The confusion matrix reveals that most instances were correctly classified, with misclassifications occurring primarily between adjacent severity levels. This outcome is clinically reasonable, as differentiating borderline cases remains challenging even for expert ophthalmologists.

Class-wise performance further supports the effectiveness of the model. Moderate DR achieved the highest scores across precision, recall, and F1-measure (>98%), suggesting that the model effectively captures the characteristic features of this stage. In contrast, slightly lower performance was observed for Mild DR and Severe NPDR (F1 ≈ 0.96), likely due to the subtle or overlapping retinal features in these stages. Nevertheless, per-class accuracies remained above 99%, demonstrating that the model does not exhibit strong bias toward any particular class.

When compared with existing studies, which commonly report classification accuracies ranging from 90% to 95% for multi-class DR tasks, the proposed approach shows significant improvement. This performance gain can be attributed to three primary factors: (i) the balanced dataset achieved through down- and up-sampling strategies, (ii) the effective representation learning capabilities of the employed model architecture, and (iii) robust training procedures that reduced overfitting.

The results highlight the model's potential for integration into clinical decision support systems. Automated DR detection and classification could substantially reduce the burden on ophthalmologists, enabling large-scale population screening and facilitating early detection, which is critical in preventing vision impairment or blindness. However, several challenges remain for real-world implementation. Variability in image acquisition quality, the presence of other retinal pathologies, and the need for generalizability across diverse patient populations may affect model performance in uncontrolled clinical environments.

Future work should address these limitations by validating the model on larger and more heterogeneous datasets, incorporating interpretability methods such as Grad-CAM to enhance clinical trust, and exploring hybrid approaches that combine deep learning with established diagnostic criteria. These advancements would further strengthen the applicability of automated DR detection systems in healthcare practice.

# 5 Conclusion

In this study, a deep learning-based approach for multi-class diabetic retinopathy (DR) classification was

developed and evaluated on a balanced dataset. The model achieved an overall accuracy of 97.21%, with consistently high precision, recall, and F1-scores across all DR stages. The confusion matrix analysis confirmed that most instances were correctly identified, with only minimal misclassifications occurring between adjacent severity levels, which are inherently difficult to distinguish even for human experts.

The results underscore the potential of the proposed method to serve as an effective tool for automated DR screening and diagnosis. By accurately identifying different stages of DR, the system can support ophthalmologists in clinical decision-making, reduce diagnostic workload, and enable large-scale population-level screening programs. Early and reliable detection is particularly significant for preventing irreversible vision loss, thereby contributing to improved patient outcomes.

Overall, this work demonstrates that deep learning holds significant promise in advancing computer-aided diagnosis for diabetic retinopathy, offering a pathway toward scalable, accurate, and accessible screening solutions.

# References


[1] WHO – Global report on diabetes.

[2] American Diabetes Association – Diabetic Retinopathy clinical facts.

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

[4] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia computer science*, *90*, 200-205.

[5] Ting, D. S., Cheung, C. Y., Nguyen, Q., Sabanayagam, C., Lim, G., Lim, Z. W., ... & Wong, T. Y. (2019). Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. *Npj Digital Medicine*, *2*(1), 24.

[6] Quellec, G., Russell, S. R., & Abràmoff, M. D. (2010). Optimal filter framework for automated, instantaneous detection of lesions in retinal images. *IEEE Transactions on medical imaging*, *30*(2), 523-533.

[7] Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *European Radiology*, *34*(12), 7895-7903.

[8] Lepetit-Aimon, G., Playout, C., Boucher, M. C., Duval, R., Brent, M. H., & Cheriet, F. (2024). MAPLES-DR: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy. *Scientific Data*, *11*(1), 914.

[9] He, C., Cao, Y., Yang, Y., Liu, Y., Liu, X., & Cao, Z. (2023). Fault diagnosis of rotating machinery based on the improved multidimensional normalization ResNet. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-11.

[10] Juola, P. (2022). Ensemble Methods. In *Encyclopedia of Big Data* (pp. 437-438). Cham: Springer International Publishing.

[11] Ghosh, D. & Chatterjee, A. (2023). Transfer-Ensemble Learning based Deep Convolutional Neural Networks for Diabetic Retinopathy Classification. Available at arXiv:2308.00525 doi: https://doi.org/10.48550/arXiv.2308.00525

[12] 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.

[13] Khan, A. Q., Sun, G., Khalid, M., Farrash, M., & Bilal, A. (2024). Multi-Deep Learning Approach With Transfer Learning for 7-Stages Diabetic Retinopathy Classification. *International Journal of Imaging Systems and Technology*, *34*(6), e23213.

[14] Hacisoftaoglu, R. E., Karakaya, M., & Sallam, A. B. (2020). Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems. *Pattern recognition letters*, *135*, 409-417.

[15] Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., & Jhanjhi, N. Z. (2022). A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. *Irbm*, *43*(4), 290-299.

[16] Jadhav, M.L., Shaikh, M.Z., Sardar, V.M. (2021). Automated Microaneurysms Detection in Fundus Images for Early Diagnosis of Diabetic Retinopathy. In: Bhateja, V., Satapathy, S.C., Travieso-González, C.M., Aradhya, V.N.M. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 1407. Springer, Singapore. https://doi.org/10.1007/978-981-16-0171-2_9

[17] Rajamani, S., & Sasikala, S. (2023). Artificial intelligence approach for diabetic retinopathy severity detection. *Informatica*, *46*(8).

[18] Zhang, Q. M., Luo, J., & Cengiz, K. (2021). An optimized deep learning-based technique for grading and extraction of diabetic retinopathy severities. *Informatica*, *45*(5).

[19] Silva, P. S., Cavallerano, J. D., Sun, J. K., & Aiello, L. M. (2021). Effectiveness of artificial intelligence–based diabetic retinopathy screening in a primary care setting: A pilot study. JAMA Ophthalmology, 139(10), 1076- 1082. doi: 10.1001/jamaophthalmol.2021.2924

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

[21] 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.

[22] 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.

[23] Zhang, G., Sun, B., Zhang, Z., Pan, J., Yang, W., & Liu, Y. (2022). Multi-model domain adaptation for diabetic retinopathy classification. *Frontiers in Physiology*, *13*, 918929.

[24] Khudhair, Z. N., Khdiar, A. N., El Abbadi, N. K., Mohamed, F., Saba, T., Alamri, F. S., & Rehman, A. (2023). Color to grayscale image conversion based on singular value decomposition. Ieee Access, 11, 54629-54638.

[25] Meng, Y., Wang, C. C., & Jin, X. (2012). Flexible shape control for automatic resizing of apparel products. Computer-aided design, 44(1), 68-76.

[26] Wang, G., Wang, Y., Bao, X., & Huang, D. (2023). Rotation has two sides: Evaluating data augmentation for deep one-class classification. In The Twelfth International Conference on Learning Representations.

[27] Pradhan, P. K., Das, A., Kumar, A., Baruah, U., Sen, B., & Ghosal, P. (2024). SwinSight: A hierarchical vision transformer using shifted windows to leverage aerial image classification. *Multimedia Tools and Applications*, *83*(39), 86457-86478.

[28] Yoo, D., Kim, J., & Yoo, J. (2024). FSwin Transformer: Feature-Space Window Attention Vision Transformer for Image Classification. *IEEE Access*, *12*, 72598-72606.

[29] Jiang, W., Cui, H., & He, K. (2024). Class-relevant Patch Embedding Selection for Few-Shot Image Classification. *arXiv preprint arXiv:2405.03722*.

[30] Lv, Y., Pan, L., Xu, K., Li, G., Zhang, W., Li, L., & Lei, L. (2025). Enhanced local multi-windows attention network for lightweight image super-resolution. *Computer Vision and Image Understanding*, *250*, 104217.

[31] Liang, Z., Zhao, K., Liang, G., Li, S., Wu, Y., & Zhou, Y. (2023). MAXFormer: Enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion. *Knowledge-Based Systems*, *280*, 110987.

[32] Lin, C., Yang, P., Wang, Q., Qiu, Z., Lv, W., & Wang, Z. (2023). Efficient and accurate compound scaling for convolutional neural networks. *Neural Networks*, *167*, 787-797.

[33] Li, Q., Luo, S., Tan, S., & Li, Z. (2025). SEAP: squeeze-and-excitation attention guided pruning for lightweight steganalysis networks. *EURASIP Journal on Information Security*, *2025*(1), 24.

[34] Zafar, A., Aamir, M., Mohd Nawi, N., Arshad, A., Riaz, S., Alruban, A., ... & Almotairi, S. (2022). A comparison of pooling methods for convolutional neural networks. *Applied Sciences*, *12*(17), 8643.