SwinNetraGF: A Gated Attention Fusion Transformer for Retinal Disease Classification

Anish Kumar Vishwakarma
School of Computer Science
UPES, Dehradun, India
anishk.vishwakarma@ddn.upes.ac.in

Preity
Department of CSE
National Institute of Technology Patna, India
preity.ph21.ec@nitp.ac.in

Rakesh Ranjan
Department of Artificial Intelligence &
AI4ICPS Innovation Hub Foundation
IIT Kharagpur, Kharagpur, India
r.ranjan@ieee.org

***Abstract*—Retinal diseases such as diabetic retinopathy and glaucoma are major causes of irreversible vision loss, highlighting the need for accurate and reliable automated diagnosis. The paper proposes *SwinNetraGF*, a gated attention fusion transformer for retinal disease classification from fundus images. The proposed framework utilizes a Swin Transformer V2 backbone to capture hierarchical multi-scale representations, enabling effective modeling of both fine-grained structural details and global contextual dependencies. To further enhance representation learning, a gated attention fusion mechanism is introduced to adaptively integrate shallow and deep features based on contextual relevance, allowing the model to emphasize diagnostically important regions while suppressing redundant information. A learnable spatial alignment module is incorporated to ensure consistency across multi-scale feature maps, facilitating efficient feature fusion. The proposed model is evaluated on a multi-class retinal dataset comprising healthy, glaucoma, and diabetic retinopathy samples under realistic clinical conditions. Experimental results demonstrate that the proposed model achieves an overall classification accuracy of 91.42% with F1-score of 0.9089 and an AUC of 0.9844. The model also achieves strong performance for diabetic retinopathy (DR) with an F1-score of 0.9795, while maintaining balanced performance for glaucoma (0.8840) and healthy (0.8634) classes despite moderate class imbalance. Qualitative analysis further demonstrates enhanced feature separability with clearly distinct clusters, validating the effectiveness of the proposed fusion strategy. These findings establish SwinNetraGF as a robust and interpretable solution for automated retinal disease classification.**

***Index Terms*—Retinal disease classification, Fundus image analysis, Gated attention mechanism, Multi-scale feature fusion.**

I. Introduction

Retinal diseases are among the leading causes of irreversible vision impairment and blindness worldwide, significantly affecting the quality of life of millions of individuals [1]. The eye conditions such as glaucoma and diabetic retinopathy (DR) are particularly critical, as they often progress silently without noticeable symptoms in the early stages [2], [3]. Fundus imaging serves as a non-invasive and widely adopted clinical tool for examining retinal structures; however, manual assessment by ophthalmologists is time-consuming, subjective, and prone to inter-observer variability. Recent advancement of artificial intelligence, particularly deep learning, automated retinal disease classification systems have gained significant attention and motivated the development of automated computer-aided diagnostic systems [4], [5].

Automated systems have demonstrated significant efficacy in the diagnosis of ocular diseases from retinal images. Several traditional convolutional neural network (CNN)-based pre-trained models, have been widely adopted for image processing tasks, particularly in the domain of ocular disease classification [6]. However, CNN-based approaches inherently depend on localized receptive fields and often struggle to capture long-range dependencies and global contextual relationships, which are essential for understanding complex retinal patterns [7]. Further, their inductive biases toward local feature extraction can limit their generalization in heterogeneous clinical environments. Recently, transformer-based architectures have emerged as powerful alternatives for visual recognition tasks due to the capability of self-attention mechanisms to capture global contextual information [8]. Among these, the Vision Transformer (ViT) has emerged as a powerful alternative to conventional CNN for image classification tasks. ViT represents images as sequences of non-overlapping patches and employs self-attention to effectively model long-range dependencies across the entire image, making it particularly suitable for medical imaging applications. Building upon the foundation of ViT, the Swin Transformer has been introduced as a hierarchical vision transformer that addresses some of the computational limitations associated with standard ViT models. By utilizing shifted window-based self-attention, the Swin Transformer [9] achieves a balance between local feature extraction and global context modeling while significantly reducing computational complexity. Its hierarchical design also allows it to generate multi-scale feature representations, which are crucial for capturing both fine-grained and coarse anatomical details in retinal images.

Recognizing the complementary capabilities of CNNs and transformers, hybrid models have been introduced to improve feature learning and classification performance. CNNs are effective in extracting low-level spatial features, while transformers excel in capturing global contextual information. Such hybrid models have demonstrated enhanced performance in fundus image classification tasks [10]–[14]. Additionally, recent works have explored feature fusion techniques and attention mechanisms to improve the discriminative capability of models by focusing on disease-relevant regions. A

few recent studies also emphasize the importance of large-scale pre-training and transfer learning in improving model generalization [5], [15]–[17]. Transformer-based models pre-trained on large datasets have shown superior robustness when adapted to medical imaging tasks, even with limited annotated data. Moreover, techniques such as data augmentation, multi-modal learning, and attention-guided frameworks have further improved classification accuracy and interpretability.

Despite these advancements, several challenges persist, including data imbalance, variability in imaging conditions, and the difficulty of detecting early-stage disease features. These challenges highlight the need for more efficient, robust, and interpretable models. Consequently, recent research is increasingly focused on integrating advanced attention mechanisms and transformer-based architectures to enhance retinal disease classification performance and enable reliable clinical deployment. Motivated by this fact we have explored Swin transformer, taking it as our baseline model, and on top of it, gated attention fusion is applied to predict retinal conditions in the human eye, such as healthy, glaucoma, and diabetic retinopathy. The key contributions of this work are summarized as follows:

- A transformer-based framework for retinal disease classification that leverages hierarchical feature extraction using Swin Transformer V2.
- A gated attention fusion mechanism that adaptively integrates shallow and deep features to enhance discriminative representation.
- A learnable spatial alignment strategy to ensure effective multi-scale feature compatibility.
- Extensive experimental evaluation demonstrating improved performance, balanced class-wise metrics, and enhanced feature separability under realistic clinical conditions.

II. Proposed Methodology

This section details the proposed hybrid architecture for retinal disease classification. The proposed approach breaks down the Swin Transformer V2 backbone used for hierarchical feature extraction and introduces a novel Gated-Attention Feature Fusion mechanism designed to selectively integrate multi-scale visual cues. Finally, we outline the token aggregation and optimization strategies that drive diagnostic performance of the proposed model.

A. Architectural Overview

We propose a hybrid transformer-based architecture optimized for the multi-class categorization of retinal fundus images (Healthy, Glaucoma, and Diabetic Retinopathy). Rather than relying on standard convolutional pipelines or isolated self-attention mechanisms, this framework leverages a Swin Transformer V2 backbone coupled with a custom Gated-Attention Feature Fusion module. The core objective is to extract and dynamically integrate multi-scale hierarchical representations, specifically, merging fine-grained spatial details from shallow layers with high-level semantic contexts from deep layers. Fig. 1 illustrates the overall architecture of the proposed framework.

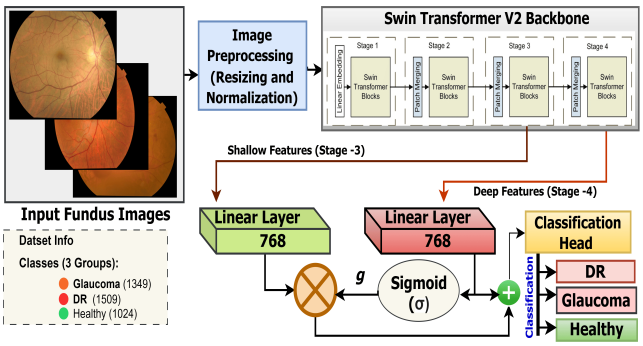


Fig. 1: Proposed framework for retinal disease classification.

B. Hierarchical Feature Extraction via Swin Transformer V2

We utilize a pre-trained Swin Transformer V2 [9] model as the primary feature extractor. Given an input retinal image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$, the backbone partitions the image into non-overlapping patches and processes them through successive hierarchical stages. We explicitly tap into the final two stages of the network to extract shallow and deep feature maps. Let these intermediate representations be denoted as $\mathbf{F}_{\text{shallow}} \in \mathbb{R}^{C_3\times \frac{H}{16}\times \frac{W}{16}}$ and $\mathbf{F}_{\text{deep}} \in \mathbb{R}^{C_4\times \frac{H}{32}\times \frac{W}{32}}$, respectively. The shallow features preserve critical micro-vascular and structural anomalies, while the deep features provide the global contextual embedding required for disease classification.

C. Spatial Downsampling and Token Alignment

To fuse $\mathbf{F}_{\text{shallow}}$ and $\mathbf{F}_{\text{deep}}$, their spatial and channel dimensions must be strictly aligned. A learnable spatial downsampling operation is applied to the shallow feature map using a strided 2×2 convolution. This reduces the spatial resolution by a factor of two while preserving local structural integrity:

$$\mathbf{F}_{\text{aligned}} = \text{Conv2D}(\mathbf{F}_{\text{shallow}}, \text{kernel} = 2, \text{stride} = 2) \quad (1)$$

where $\mathbf{F}_{\text{aligned}} \in \mathbb{R}^{C_3\times \frac{H}{32}\times \frac{W}{32}}$.

Both feature maps are then flattened along their spatial dimensions and transposed to generate sequence tokens, $\mathbf{T}_{\text{shallow}} \in \mathbb{R}^{N\times C_3}$ and $\mathbf{T}_{\text{deep}} \in \mathbb{R}^{N\times C_4}$, where the sequence length is $N = \frac{H}{32}\times\frac{W}{32}$.

To map the shallow tokens into the identical embedding dimension D (where $D = 768$) as the deep tokens, we apply a linear projection matrix $\mathbf{W}_{p3}$:

$$\hat{\mathbf{T}}_{\text{shallow}} = \mathbf{T}_{\text{shallow}}\mathbf{W}_{p3} \quad (2)$$

D. Gated-Attention Feature Fusion Mechanism

A naive addition or concatenation of $\hat{\mathbf{T}}_{\text{shallow}}$ and $\mathbf{T}_{\text{deep}}$ often introduces representational noise, as not all shallow spatial details are relevant to the global context. To address this, we introduce a Gated-Attention Feature Fusion module that adaptively modulates the flow of shallow features based on the semantic state of the deep features.

First, we compute a gating signal $\mathbf{G} \in \mathbb{R}^{N \times D}$ entirely derived from the deep representation. This is achieved through a learnable linear transformation $\mathbf{W}_g$, followed by layer normalization and a sigmoid activation function σ:

$$\mathbf{G} = \sigma\left(\text{LayerNorm}(\mathbf{T}_{\text{deep}}\mathbf{W}_g)\right) \tag{3}$$

The matrix $\mathbf{G}$ acts as a soft-attention mask, with values bounded in $(0, 1)$, representing the contextual importance of each token dimension. The shallow tokens are subsequently passed through a secondary projection $\mathbf{W}_f$ and modulated element-wise by $\mathbf{G}$. The final fused representation $\mathbf{F}_{\text{fused}}$ is computed via a residual connection with the deep tokens:

$$\mathbf{F}_{\text{fused}} = \mathbf{T}_{\text{deep}} + \left(\mathbf{G} \odot (\hat{\mathbf{T}}_{\text{shallow}}\mathbf{W}_f)\right) \tag{4}$$

where $\odot$ denotes the Hadamard product. It practically achieves a selective emphasis pipeline: the model retains the rich semantic floor of $\mathbf{T}_{\text{deep}}$ while only activating the high-resolution features of $\hat{\mathbf{T}}_{\text{shallow}}$ that explicitly contribute to the diagnostic task.

To stabilize the variance of the fused tokens prior to classification, a final layer normalization is applied:

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{F}_{\text{fused}}) \tag{5}$$

E. Classification Head and Optimization Strategy

The sequence of fused tokens $\mathbf{Z} \in \mathbb{R}^{N \times D}$ is aggregated into a single global representation vector $\mathbf{z}_{\text{pool}} \in \mathbb{R}^D$ via Global Average Pooling along the token dimension:

$$\mathbf{z}_{\text{pool}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i \tag{6}$$

This vector is routed through the classification head, which comprises Layer Normalization, Dropout ($p = 0.5$) to mitigate overfitting, and a final linear layer to project the embeddings into the $K = 3$ class logits:

$$\hat{\mathbf{y}} = \mathbf{W}_c\left(\text{Dropout}(\text{LayerNorm}(\mathbf{z}_{\text{pool}}))\right) + \mathbf{b}_c \tag{7}$$

Optimization and Loss: To penalize overconfidence in the model's predictions, we train the network using a Label Smoothing Cross-Entropy loss function. For a ground-truth label y and predicted class probabilities p_k, the loss is formulated as:

$$\mathcal{L}_{LS} = -(1-\alpha)\log(p_y) - \frac{\alpha}{K}\sum_{k=1}^{K}\log(p_k) \tag{8}$$

where $\alpha = 0.15$ is the smoothing parameter.

The network is optimized using AdamW with a weight decay of 0.05. We utilize a differential learning rate strategy to preserve the pre-trained knowledge of the backbone while aggressively updating the newly initialized fusion and classification layers. Specifically, the Swin V2 backbone is updated with a base learning rate of 1×10^{-5}, whereas the downstream parameters are updated at 5×10^{-4}. The learning rates are annealed using a Cosine Annealing schedule over the training duration. Additionally, mixed-precision training and gradient clipping ($\max \|\nabla\|_2 = 1.0$) are enforced to ensure numerical stability and computational efficiency.

Algorithm 1 Training Procedure for the Proposed Framework

Require: Dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, epochs E, learning rates $(\eta_{\text{base}}, \eta_{\text{head}})$, label smoothing α
Ensure: Optimized parameters Θ
Initialize Swin-V2 backbone f_{swin} and parameters $\{\mathbf{W}_{p3}, \mathbf{W}_g, \mathbf{W}_f, \mathbf{W}_c, \mathbf{b}_c\}$
Initialize AdamW optimizer with decoupled learning rates and cosine annealing scheduler
for $e = 1$ to E **do**
  for mini-batch $(\mathbf{X}, y) \subset \mathcal{D}$ **do**
    $\mathbf{F}_{\text{shallow}}, \mathbf{F}_{\text{deep}} \leftarrow f_{\text{swin}}(\mathbf{X})$
    $\hat{\mathbf{T}}_{\text{shallow}} \leftarrow \text{Flatten}(\text{Conv2D}_{\downarrow 2}(\mathbf{F}_{\text{shallow}}))^\top \mathbf{W}_{p3}$
    $\mathbf{T}_{\text{deep}} \leftarrow \text{Flatten}(\mathbf{F}_{\text{deep}})^\top$
    $\mathbf{G} \leftarrow \sigma(\text{LayerNorm}(\mathbf{T}_{\text{deep}}\mathbf{W}_g))$
    $\mathbf{Z} \leftarrow \text{LayerNorm}(\mathbf{T}_{\text{deep}} + \mathbf{G} \odot (\hat{\mathbf{T}}_{\text{shallow}}\mathbf{W}_f))$
    $\hat{\mathbf{y}} \leftarrow \mathbf{W}_c(\text{Dropout}(\text{GAP}(\mathbf{Z}))) + \mathbf{b}_c$
    $\mathcal{L} \leftarrow \mathcal{L}_{LS}(\hat{\mathbf{y}}, y, \alpha)$
    Compute $\nabla_\Theta \mathcal{L}$, gradient clipping $(\|\nabla_\Theta\|_2 \leq 1)$
    Update backbone with η_{base} and head with η_{head}
  end for
  Update learning rate via cosine annealing
end for
return Θ

III. Experiments and Results

This section presents the experimental setup and evaluates the performance of the proposed model on the retinal image classification task. The dataset and training configuration are first described. Subsequently, quantitative and qualitative results are reported to demonstrate the effectiveness of the proposed approach.

A. Dataset and Implementation Details

1) Dataset Description: The proposed framework is evaluated using the Eye Disease Image Dataset available at Mendeley Data repository [18], which contains 5,335 color fundus images collected from real-world clinical environments in Bangladesh. Originally, the dataset comprises ten diagnostic categories; however, this study focuses exclusively on primary retinal conditions. Accordingly, three clinically significant classes are considered: *DR* (1,509 images), *Glaucoma* (1,349 images), and *Healthy* fundus scans (1,024 images). This selection results in a curated subset of 3,882 images used for model development and evaluation. From a statistical perspective, the dataset exhibits moderate class imbalance, which can potentially bias the learning process toward majority classes if not properly addressed. To mitigate this issue without altering the inherent data distribution, a weighted random sampling strategy is employed. For robust and unbiased evaluation, the dataset is partitioned into three mutually exclusive subsets following a standard protocol: 70% for training, 15% for validation, and 15% for testing.

2) Implementation and Training Protocols: The proposed model is implemented using the *PyTorch* framework. During training, input images undergo stochastic augmentation through random resizing and cropping to 256×256. During evaluation, images are resized to 288 pixels and center-cropped

TABLE I: Performance evaluation of the proposed model for retinal disease classification.

Class	Accuracy	Precision	Recall	Specificity	F1-score	AUC
DR	0.9862	0.9729	0.9862	0.9836	0.9795	0.9993
Glaucoma	0.8818	0.8861	0.8818	0.9395	0.8840	0.9765
Healthy	0.8580	0.8688	0.8580	0.9501	0.8634	0.9775
Overall	0.9142	0.9092	0.9087	0.9577	0.9089	0.9844

to 256×256. All inputs are normalized using standard ImageNet channel statistics. A differential learning rate strategy is adopted using the AdamW optimizer to effectively train the hybrid architecture. The pretrained Swin V2 backbone is fine-tuned with a learning rate of 1×10^{-5}, while the gated fusion and classification layers are optimized with a higher learning rate of 5×10^{-4}. The network is trained for 40 epochs using cosine annealing scheduling with a weight decay of 0.05. To improve robustness and mitigate overconfidence, label smoothing with $\alpha = 0.15$ is applied to the cross-entropy loss. Training is performed with a batch size of 16 using mixed-precision computation on CUDA-enabled GPUs. Additionally, gradient clipping with a maximum norm of 1.0 is employed to ensure stable optimization. The overall training strategy of the proposed framework is summarized in Algorithm 1.

B. Quantitative Analysis

This section presents a comprehensive evaluation of the proposed framework using standard classification and discriminative metrics. The analysis includes accuracy, class-wise performance, confusion matrix interpretation, and convergence behavior. As depicted in Table I, the proposed model achieves an overall classification accuracy of 91.42%, indicating reliable performance across all diagnostic categories. In particular, DR is identified with high precision (0.9729), recall (0.9862), and F1-score (0.9795), while the macro-averaged F1-score of 0.9089 reflects balanced performance despite moderate class imbalance. The proposed model demonstrates improved performance compared to the ensemble transfer learning approaches. For the relevant classes, the Ensemble TL model [19] reports F1-scores of 0.92 (DR), 0.66 (Glaucoma), and 0.76 (Healthy), while the Ensemble TF model [19] achieves 0.92, 0.66, and 0.78, respectively. In comparison, the proposed method attains higher F1-scores of 0.9795 (DR), 0.8840 (Glaucoma), and 0.8634 (Healthy), along with a macro AUC of 0.9844, indicating superior classification performance for these classes.

Confusion matrix in Fig. 2 shows that DR samples are not misclassified as Healthy or Glaucoma, with only negligible confusion observed. However, misclassification is more prominent between Glaucoma and Healthy classes, where 31 Glaucoma images are predicted as Healthy and 20 Healthy images as Glaucoma. This can be attributed to the distinct visual features of DR (e.g., hemorrhages and exudates), whereas Glaucoma exhibits subtle structural variations such as optic cup-to-disc ratio changes that overlap with normal anatomy. Nevertheless, the model maintains strong F1-scores of 0.8840 and 0.8634 for Glaucoma and Healthy classes, respectively.

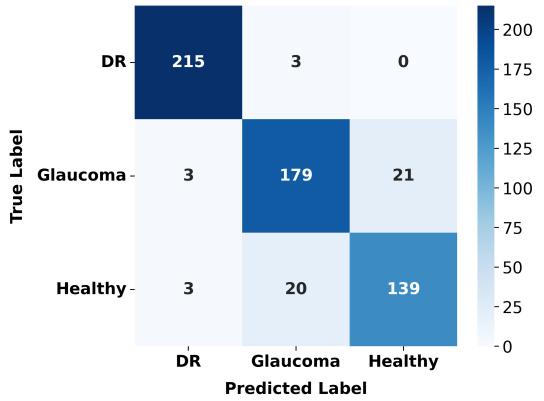
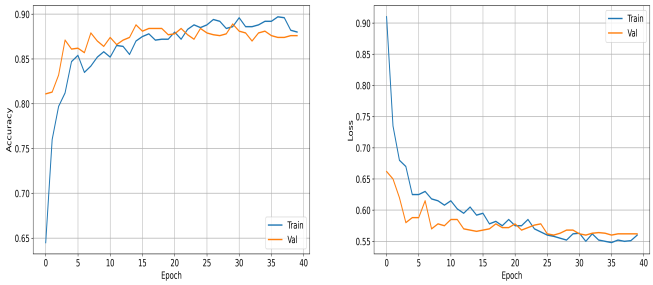


Fig. 2: Confusion matrix illustrating the classification performance of the proposed model for three classes: Diabetic Retinopathy (DR), Glaucoma, and Healthy.



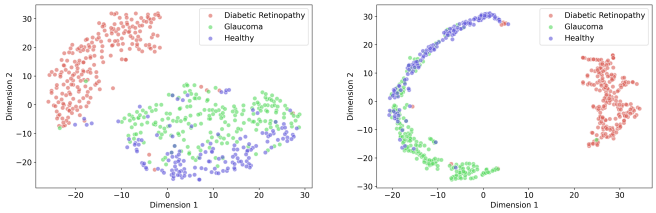
(a) Accuracy curves. (b) Loss curves.

Fig. 3: Training and validation accuracy and loss curves of the proposed model across epochs.

(a) Healthy (b) Glaucoma (c) DR

Fig. 4: Sample images showing correct classification by the proposed model for three classes.

The accuracy trends further demonstrate stable convergence, with validation accuracy closely tracking training accuracy. This behavior indicates effective generalization, supported by label smoothing ($\alpha = 0.15$), dropout, and the adopted optimization strategy. Additionally, the training dynamics (Fig. 3a, Fig. 3b) confirm stable optimization, with both training and validation losses converging to 0.55 after epoch 40 without divergence.



(a) After the backbone network. (b) After the classifier head.

Fig. 5: Visualization of feature representations for Diabetic Retinopathy (DR), Glaucoma, and Healthy samples.

C. Qualitative Analysis

To further understand the effectiveness of the proposed model, a qualitative analysis is conducted by examining prediction outputs and learned feature representations, as illustrated in Fig. 4 and Fig. 5. Fig. 4 presents representative images from the three classes: *Healthy*, *Glaucoma*, and *DR*. The predicted labels match the ground truth with high confidence scores, indicating that the proposed model effectively captures discriminative retinal features and accurately distinguishes between normal and diseased conditions.

To further analyze the learned feature space, Fig. 5 visualizes the embeddings at two stages of the network. Fig. 5(a) shows the feature distribution obtained after the *backbone network*, where the class clusters begin to form but still exhibit slight overlap. In contrast, Fig. 5(b) presents the embeddings after the *classifier head*, where the clusters become more compact and clearly separated. This improved separation demonstrates that the classifier further refines the feature representations learned by the backbone, enhancing the model's discriminative capability. Overall, the qualitative results confirm that the proposed framework learns meaningful and well-structured feature representations, leading to reliable multi-class retinal disease classification.

IV. Conclusion

The proposed method utilizes a transformer-based framework with gated attention fusion to effectively integrate multi-scale retinal features for robust disease classification. By combining hierarchical representations from the Swin Transformer with adaptive feature modulation, the model enhances discriminative capability while reducing inter-class ambiguity, particularly in clinically challenging cases. The results demonstrate consistent and balanced performance across all categories, supported by improved feature separability and stable generalization behavior under realistic data conditions. The proposed design highlights the importance of selective multi-scale feature integration within transformer architectures for medical imaging tasks. Future work will focus on extending the framework by incorporating additional disease categories and severity grading. Optimizing the model for real-time clinical deployment could be one of the potential future goals.

References

[1] M. H. Sarhan, M. A. Nasseri, D. Zapp, M. Maier, C. P. Lohmann, N. Navab, and A. Eslami, "Machine learning techniques for ophthalmic data processing: a review," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3338–3350, 2020.

[2] H. Datta, Preity, and A. K. Bhandari, "Recent trends for computational enriched diabetic retinopathy assessment: A systematic review," *Archives of Computational Methods in Engineering*, vol. 33, no. 1, pp. 1227–1256, 2026.

[3] A. Bali and V. Mansotra, "Analysis of deep learning techniques for prediction of eye diseases: A systematic review," *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 487–520, 2024.

[4] Preity, A. K. Bhandari, and S. Shahnawazuddin, "Automated computationally intelligent methods for ocular vessel segmentation and disease detection: a review," *Archives of Computational Methods in Engineering*, vol. 31, no. 2, pp. 701–724, 2024.

[5] M. Juneja, S. Thakur, A. Uniyal, A. Wani, N. Thakur, and P. Jindal, "Deep learning-based classification network for glaucoma in retinal images," *Computers and Electrical Engineering*, vol. 101, p. 108009, 2022.

[6] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[7] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[8] Z. Wang, Y. Chen, Y. Gu, J. Liu, X. Zhu, and M. He, "The evolution of object detection from cnns to transformers and multi-modal fusion," *Scientific Reports*, vol. 16, no. 1, p. 7517, 2026.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows," 2021.

[10] A. M. Towhid, F. Yasmin, and S. Hasan, "Eyelinenet: An ensemble deep learning for eye health monitoring with attention mechanisms and grad-cam visualization," in *2025 Eighth International Conference on Image Information Processing (ICIIP)*, 2025, pp. 705–710.

[11] P. Srikanth, Y. C. A. Padmanabha Reddy, and K. Kishan Babu, "Seefusenet: A stage-enhanced multi-level feature fusion framework for eye disease classification," in *2025 IEEE 22nd India Council International Conference (INDICON)*, 2025, pp. 1–6.

[12] A. K. M. Masum, M. F. I. Khan, M. M. Hassan, D. M. Farid, A. K. Bitto, and M. A. Rahman, "Multi-model ensemble approach for accurate classification of ocular disorders," in *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, 2025, pp. 1–6.

[13] Y. Liu, D. Yao, Y. Ma, H. Wang, J. Wang, X. Bai, G. Zeng, and Y. Liu, "Stmf-drnet: A multi-branch fine-grained classification model for diabetic retinopathy using swin-transformerv2," *Biomedical Signal Processing and Control*, vol. 103, p. 107352, 2025.

[14] S. Sweidan and A. Taha, "Deepretina framework for multi-retinal diseases classification," *Arabian Journal for Science and Engineering*, 2025. [Online]. Available: https://doi.org/10.1007/s13369-025-10888-2

[15] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi, "Retinal image synthesis and semi-supervised learning for glaucoma assessment," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2211–2218, 2019.

[16] J. Wang, Y. Bai, and B. Xia, "Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3397–3407, 2020.

[17] D. Das, D. R. Nayak, and R. B. Pachori, "Ca-net: A novel cascaded attention-based network for multistage glaucoma classification using fundus images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[18] R. Rashid, S. Sharmin, T. Khatun, M. Z. Hasan, and M. S. Uddin, "Eye disease image dataset," 2024. [Online]. Available: https://doi.org/10.17632/s9bfhswzjb.1

[19] A. K. M. Masum, M. F. I. Khan, M. M. Hassan, D. M. Farid, A. K. Bitto, and M. A. Rahman, "Multi-model ensemble approach for accurate classification of ocular disorders," in *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*. IEEE, 2025, pp. 1–6.