

# اسراء الوتار

## An Efficient Hybrid Framework Combining Transformer Feature Encoding and Machine Learning for Diabetic Retinopa...

 Midocean University

### Document Details

**Submission ID**

trn:oid:::2945:406895992

**Submission Date**

Aug 13, 2026, 10:25 AM GMT+3

**Download Date**

Aug 13, 2026, 2:30 PM GMT+3

**File Name**

An Efficient Hybrid Framework Combining Transformer Feature Encoding and Machine Learnin....docx

**File Size**

7.9 MB

56 Pages

8,759 Words

54,448 Characters

## 36% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



40 AI-generated 36%

Likely generated or likely generated and revised by AI.

### Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



# **An Efficient Hybrid Framework Combining Transformer Feature Encoding and Machine Learning for Diabetic Retinopathy Classification**

Prepared By

Name: Asraa Alwattar

ID No.: 253001066

Name: Rahaf Barakat

ID No.: 253002370

Submitted in fulfilment of the requirements for the degree Master of Science in  
Artificial Intelligence

At

College of Informatics  
Midocean University  
UAE

Supervisor Name: Wessam Ahmad

College of Informatics, Midocean University, Comoros

Email:.....

## Acknowledgements

Special thanks to Dr. .

## Abstract

Diabetic retinopathy (DR) is a major cause of avoidable visual impairment, yet grading its severity from fundus photographs is difficult because image quality varies, disease classes are imbalanced, and neighboring stages often share similar visual signs. This thesis evaluates a five-class grading pipeline that combines deep feature extraction with a conventional machine-learning classifier. Two pretrained backbones were fine-tuned on the APTOS 2019 Blindness Detection dataset: ConvNeXt-Tiny to capture local and hierarchical retinal structure, and Swin-V2-Tiny to add broader contextual information. The dataset contains 3,662 images labeled as No DR, Mild, Moderate, Severe, or Proliferative DR. A 768-dimensional embedding was obtained from each backbone, the two vectors were concatenated, and a regularized Random Forest was trained for the final prediction. On the held-out test set, the framework achieved 84.91% accuracy, 73.93% macro precision, 68.99% macro recall, 70.98% macro F1-score, 96.11% macro specificity, and 94.80% macro ROC-AUC. The results indicate that combining the two feature families provides useful complementary information, while the class-level results also show that minority severity grades remain harder to identify reliably. These findings support further work on external validation, class-sensitive learning, and improved recognition of underrepresented stages.

**Keywords:** Diabetic Retinopathy; Deep Learning; ConvNeXt-Tiny; Swin-V2-Tiny; Random Forest; Multiclass Classification; Medical Image Analysis.

## List of Publications

Alwattar, A., & Barakat, R. *A Comprehensive Review of Artificial Intelligence Approaches for Diabetic Retinopathy Classification*. Manuscript in preparation, College of Informatics, Midocean University.

# Table of Contents

|                                                                    |           |
|--------------------------------------------------------------------|-----------|
| <b>Acknowledgements .....</b>                                      | <b>2</b>  |
| <b>Abstract.....</b>                                               | <b>3</b>  |
| <b>List of Publications .....</b>                                  | <b>3</b>  |
| <b>Table of Contents .....</b>                                     | <b>5</b>  |
| <b>List of Abbreviations .....</b>                                 | <b>7</b>  |
| <b>List of Figures.....</b>                                        | <b>8</b>  |
| <b>List of Tables .....</b>                                        | <b>9</b>  |
| <b>Chapter 1 Introduction.....</b>                                 | <b>10</b> |
| 1.1 Thesis Overview .....                                          | 10        |
| 1.2 Problem Definition .....                                       | 11        |
| 1.3 Thesis Objectives .....                                        | 11        |
| 1.4 Thesis Contribution .....                                      | 12        |
| 1.5 Thesis Organization.....                                       | 13        |
| <b>Chapter 2: Background and Related Work .....</b>                | <b>14</b> |
| 2.1 Introduction .....                                             | 14        |
| 2.2 Background .....                                               | 14        |
| 2.2.1 Diabetic Retinopathy Overview.....                           | 14        |
| 2.2.2 Artificial Intelligence in Medical Image Analysis:.....      | 15        |
| 2.2.3 Diabetic Retinopathy Detection Using Machine Learning .....  | 16        |
| 2.2.4 Diabetic Retinopathy Detection Using Deep Learning.....      | 19        |
| 2.2.5 Diabetic Retinopathy Detection Using Transformer Models..... | 20        |
| 2.2.6 Deep Feature Fusion and Hybrid Classification .....          | 23        |
| 2.2.7 Performance Metrics.....                                     | 24        |
| 2.3 Literature Review .....                                        | 27        |
| 2.4 Comparative Analysis of Related Work.....                      | 31        |
| 2.5 Chapter Summary .....                                          | 33        |
| <b>Chapter 3: Proposed Approach (Methodology).....</b>             | <b>34</b> |
| 3.1 Introduction .....                                             | 34        |
| 3.2 Proposed Framework.....                                        | 34        |
| 3.3 Dataset Description .....                                      | 35        |
| 3.4 Data Preprocessing .....                                       | 37        |
| 3.5 Dataset Splitting and Training-Only Augmentation.....          | 38        |
| 3.5.1 Stratified Dataset Partitioning.....                         | 38        |
| 3.5.2 Training-Only Data Augmentation .....                        | 39        |
| 3.6 Deep Feature Extraction .....                                  | 39        |
| 3.6.1 ConvNeXt-Tiny .....                                          | 41        |
| 3.6.2 Swin-V2-Tiny .....                                           | 40        |
| 3.7 Feature Fusion .....                                           | 40        |
| 3.8 Random Forest Classification.....                              | 41        |

|                                                             |           |
|-------------------------------------------------------------|-----------|
| 3.9 Chapter Summary .....                                   | 42        |
| <b>Chapter 4: Experimental Results and Discussion.....</b>  | <b>43</b> |
| 4.1 Introduction .....                                      | 43        |
| 4.2 Experimental Setup .....                                | 43        |
| 4.3 Results of the Final Hybrid Model .....                 | 44        |
| 4.4 Discussion of Findings .....                            | 49        |
| 4.4.1 Contribution of Feature Fusion and Random Forest..... | 49        |
| 4.4.2 Interpretation in Relation to Related Work.....       | 49        |
| 4.5 Limitations.....                                        | 50        |
| 4.6 Chapter Summary .....                                   | 51        |
| <b>Chapter 5 Conclusion and Future Work.....</b>            | <b>52</b> |
| <b>References .....</b>                                     | <b>54</b> |

## List of Abbreviations

- i. Diabetic Retinopathy (DR)
- ii. Machine Learning (ML)
- iii. Deep Learning (DL)
- iv. Convolutional Neural Networks (CNNs)
- v. Support Vector Machines (SVM)
- vi. True Positives (TP)
- vii. True Negatives (TN)
- viii. False Positives (FP)
- ix. False Negatives (FN)
- x. Random Forest (RF)
- xi. Vision Transformer (ViT)
- xii. Asia Pacific Tele-Ophthalmology Society (APTOS)
- xiii. Contrast Limited Adaptive Histogram Equalization (CLAHE)
- xiv. Receiver Operating Characteristic (ROC)
- xv. Area Under the Curve (AUC)
- xvi. Graphics Processing Unit (GPU)
- xvii. Quadratic Weighted Kappa (QWK)

## List of Figures:

|                                                                                             |    |
|---------------------------------------------------------------------------------------------|----|
| Figure 1: AI Techniques for Retinal Image Analysis                                          | 16 |
| Figure 2: Support Vector Machine Classification of Retinal Features                         | 17 |
| Figure 3: Random Forest Classification for Diabetic Retinopathy Grading                     | 18 |
| Figure 4: K-Nearest Neighbors for Diabetic Retinopathy Grading                              | 19 |
| Figure 5: ConvNeXt-Tiny Feature Extraction for Diabetic Retinopathy Grading                 | 20 |
| Figure 6: Transformer-Based Retinal Image Encoding                                          | 21 |
| Figure 7: Vision Transformer Encoding for Diabetic Retinopathy Analysis                     | 22 |
| Figure 8: Swin Transformer Encoding for Diabetic Retinopathy Grading                        | 23 |
| Figure 9: ConvNeXt-Swin Feature Fusion and Random Forest Classification                     | 24 |
| Figure 10: Proposed Hybrid Framework for Diabetic Retinopathy Classification                | 35 |
| Figure 11: Original Class Distribution in the APTOS 2019 Dataset                            | 36 |
| Figure 12: Representative Retinal Images Before Preprocessing                               | 36 |
| Figure 13: Representative Retinal Images After Preprocessing                                | 38 |
| Figure 14: Confusion Matrix of the Final Hybrid Model                                       | 45 |
| Figure 15: One-vs-Rest ROC Curves of the Final Hybrid Model                                 | 46 |
| Figure 16: Bar Chart of the Comprehensive Performance Metrics of the Final Hybrid Framework | 47 |
| Figure 17: Bar Chart Comparison of the Proposed Framework with Selected Related Studies     | 48 |

## List of Tables:

|                                                                             |    |
|-----------------------------------------------------------------------------|----|
| Table 1: Comparative study of diabetic retinopathy detection methods        | 32 |
| Table 2: Deterministic Retinal-Image Preprocessing Parameters               | 37 |
| Table 3: ConvNeXt-Tiny Final Parameters                                     | 39 |
| Table 4: Swin-V2-Tiny Final Parameters                                      | 40 |
| Table 5: Selected Random Forest Parameters                                  | 42 |
| Table 6: Experimental Dataset Split                                         | 44 |
| Table 7: Comprehensive Performance of the Final Hybrid Framework            | 44 |
| Table 8: Comparison of the Proposed Framework with Selected Related Studies | 48 |

# Chapter 1 Introduction

## 1.1 Thesis Overview

Diabetic retinopathy is a retinal complication of long-term diabetes in which persistent hyperglycemia damages the small retinal vessels. As the damage progresses, fundus images may show microaneurysms, hemorrhages, exudates, ischemic changes, and eventually neovascularization [1]. In routine practice, clinicians grade disease severity from retinal photographs, but this process depends on specialist review and becomes difficult to scale when screening large populations [2]. These limitations have encouraged the use of AI methods that learn diagnostic patterns directly from fundus images [3]. The present study addresses five severity levels by combining fixed image preprocessing, augmentation restricted to the training set, two pretrained backbones (ConvNeXt-Tiny and Swin-V2-Tiny), feature concatenation, and a Random Forest classifier. The two backbones were selected because they emphasize different visual information: ConvNeXt focuses strongly on local structure, while Swin-V2-Tiny also models relationships across wider retinal regions.

## 1.2 Problem Definition

Five-class diabetic retinopathy grading is difficult for several practical reasons. Fundus images do not arrive with uniform resolution, illumination, contrast, field of view, or acquisition quality, and the number of examples is uneven across severity levels. Distinguishing adjacent grades is especially challenging because the visual transition from one stage to the next is gradual, while Severe and Proliferative DR are represented by fewer images. A CNN can capture localized lesion and texture information well, but a single convolutional representation may not describe broader spatial context sufficiently. Transformer models provide a different view by relating information across retinal regions, although they can be sensitive to data volume and regularization. For this reason, the study examines whether ConvNeXt-Tiny and Swin-V2-Tiny embeddings can be combined and then classified with a regularized Random Forest. The experimental procedure also uses integrity checks, stratified data splitting, class weighting, training-only augmentation, a held-out validation set, and a final untouched test set so that model selection and evaluation remain separate.

## 1.3 Thesis Objectives

The main goal of this thesis is to improve the hybrid approach towards diagnosis of diabetic retinopathy in five classes using retinal fundus images. In the proposed research, attention is given to the pre-processing and normalization of the APTOS 2019 data set. Imbalance in the data set is addressed through the application of stratified split and class balancing, augmentation of data only during training, tuning of two feature extractors such as ConvNeXt-Tiny and Swin-V2-Tiny, obtaining a 768D embedding from both the models, and getting a hybrid 1,536D embedding by comparing predefined candidate configurations on the held-out validation subset and evaluate the final framework using accuracy, precision, recall, F1-score, specificity,

confusion matrices, and one-vs-rest AUC measures. The performance of the hybrid classifier is subsequently interpreted in relation to the reviewed literature.

## 1.4 Thesis Contribution

This thesis contributes a reproducibility-oriented hybrid methodology for diabetic retinopathy severity grading through the following elements:

1. It proposes a novel hybrid visual representation that fuses convolutional and transformer-based feature extraction by fine-tuning ConvNeXt-Tiny and Swin-V2-Tiny in parallel capturing both fine-grained local lesion patterns and global contextual spatial relationships within retinal images.
2. Implements a regularized Random Forest as the final multiclass classifier and selects its configuration through held-out validation comparison without using the testing subset for model selection.
3. Evaluates the proposed hybrid classifier using class-sensitive measures, including per-class and aggregate precision, recall, F1-score, and specificity; confusion-matrix analysis; and macro and weighted one-vs-rest ROC-AUC, with particular attention to imbalanced severity grades.

## 1.5 Thesis Organization

The rest of this dissertation will be organized as follows:

- **Chapter 2** presents the medical and technical background of diabetic retinopathy grading. It reviews image preprocessing, data augmentation, machine learning, convolutional networks, vision transformers, hybrid feature representations, evaluation measures, and fourteen related studies that motivate the proposed methodology.
- **Chapter 3** describes the proposed methodology in detail, including APTOS 2019 dataset validation, deterministic preprocessing, stratified splitting, training-only augmentation, ConvNeXt-Tiny and Swin-V2-Tiny fine-tuning, deep-embedding extraction, feature fusion, Random Forest model selection, and the evaluation protocol.
- **Chapter 4** presents the experimental findings, including overall and class-level results, and discusses how the final hybrid model behaved on the held-out data.
- **Chapter 5** closes the thesis by summarizing the main findings, outlining the study's limitations, and identifying practical directions for future work such as explainability and external clinical validation.

## Chapter 2: Background and Related Work

### 2.1 Introduction

This chapter provides the background needed to understand the proposed approach. It first reviews diabetic retinopathy and the clinical meaning of its severity grades. The discussion then turns to the main computational components used in this study: preprocessing, augmentation, transfer learning, conventional classifiers, convolutional networks, vision transformers, deep embeddings, and feature fusion. Because the dataset is imbalanced, the chapter also explains the evaluation measures used later in the experiments. The final part reviews fourteen related studies and uses them to position the ConvNeXt-Tiny, Swin-V2-Tiny, and Random Forest combination adopted here.

### 2.2 Background

#### 2.2.1 Diabetic Retinopathy Overview

Diabetic retinopathy develops when prolonged diabetes damages retinal microvessels. On fundus examination, this damage may appear as microaneurysms, hemorrhages, hard exudates, cotton-wool spots, venous abnormalities, intraretinal microvascular abnormalities, or neovascularization. Clinically, the condition is graded from no DR through mild, moderate, and severe non-proliferative disease to proliferative DR. The grade matters because both the risk of vision-threatening complications and the urgency of treatment increase as disease severity rises. Since early retinal changes can occur before a patient notices symptoms, regular fundus screening is important for timely detection [4].

### 2.2.2 Artificial Intelligence in Medical Image Analysis:

Artificial intelligence is now widely used in medical image analysis because learning-based models can extract visual patterns that are difficult to specify manually. In retinal applications, an automated workflow commonly includes quality checking, preprocessing, augmentation, representation learning, classification, and evaluation.

Preprocessing reduces irrelevant variation by applying operations such as cropping, contrast adjustment, resizing, and normalization. Augmentation is used during training to expose the model to additional variations of the available images, which can be helpful when the class distribution is uneven. Traditional machine-learning pipelines usually separate feature extraction from classification, whereas deep networks learn hierarchical representations directly from images. Transfer learning is particularly useful for retinal datasets because a model pretrained on a larger image collection can be adapted to the target task instead of being learned from scratch. Hybrid approaches extend this idea by extracting embeddings from more than one backbone and combining them before classification. In this thesis, convolutional features are used for local retinal structure and transformer features for broader context, which motivates the fusion strategy [4], [5]. Figure 1 summarizes the main AI approaches discussed in this section.

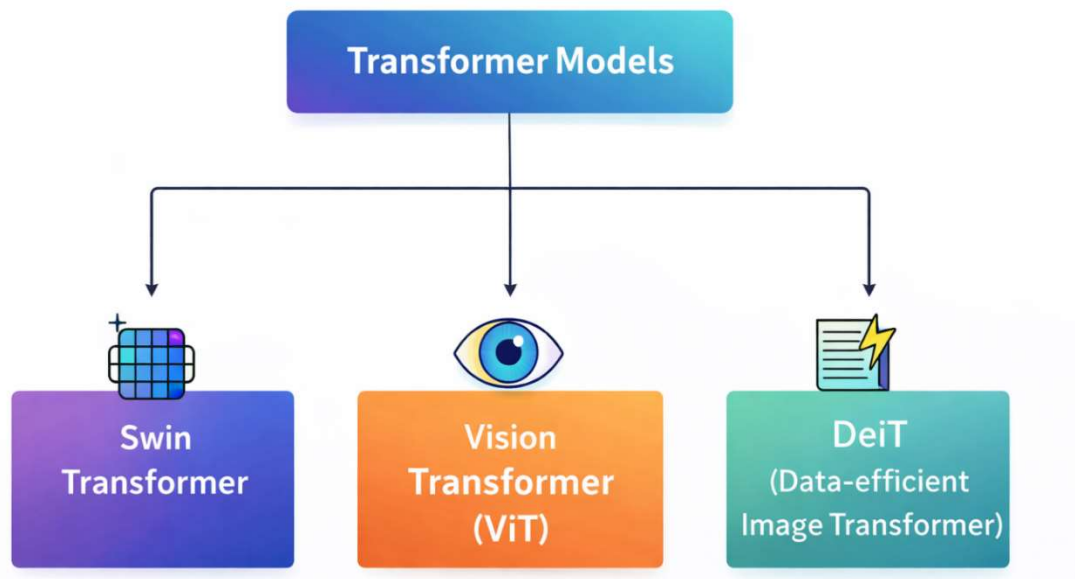


Figure 1: AI Techniques for Retinal Image Analysis [4] [5]

### 2.2.3 Diabetic Retinopathy Detection Using Machine Learning

In a conventional machine-learning pipeline, feature construction and classification are treated as separate stages. A retinal image can first be represented by handcrafted descriptors of texture, intensity, vessels, or shape, or by features produced by a pretrained neural network. A second model then performs the actual class assignment. This separation is useful when several classifiers need to be tested on the same representation because the deep feature extractor does not have to be retrained each time. The proposed framework follows this idea: ConvNeXt-Tiny and Swin-V2-Tiny generate the image embeddings, the two embeddings are fused, and Random Forest performs the final five-class prediction.

#### 1. Support Vector Machine (SVM)

A Support Vector Machine searches for a decision boundary with the largest possible margin between classes. Kernel functions can be introduced when the relationship between classes is not linearly separable. SVMs have been used with both engineered and learned retinal features, but their behavior can change with feature

scaling, kernel selection, and hyperparameter settings [6]. Figure 2 gives a simple illustration of this classification principle.

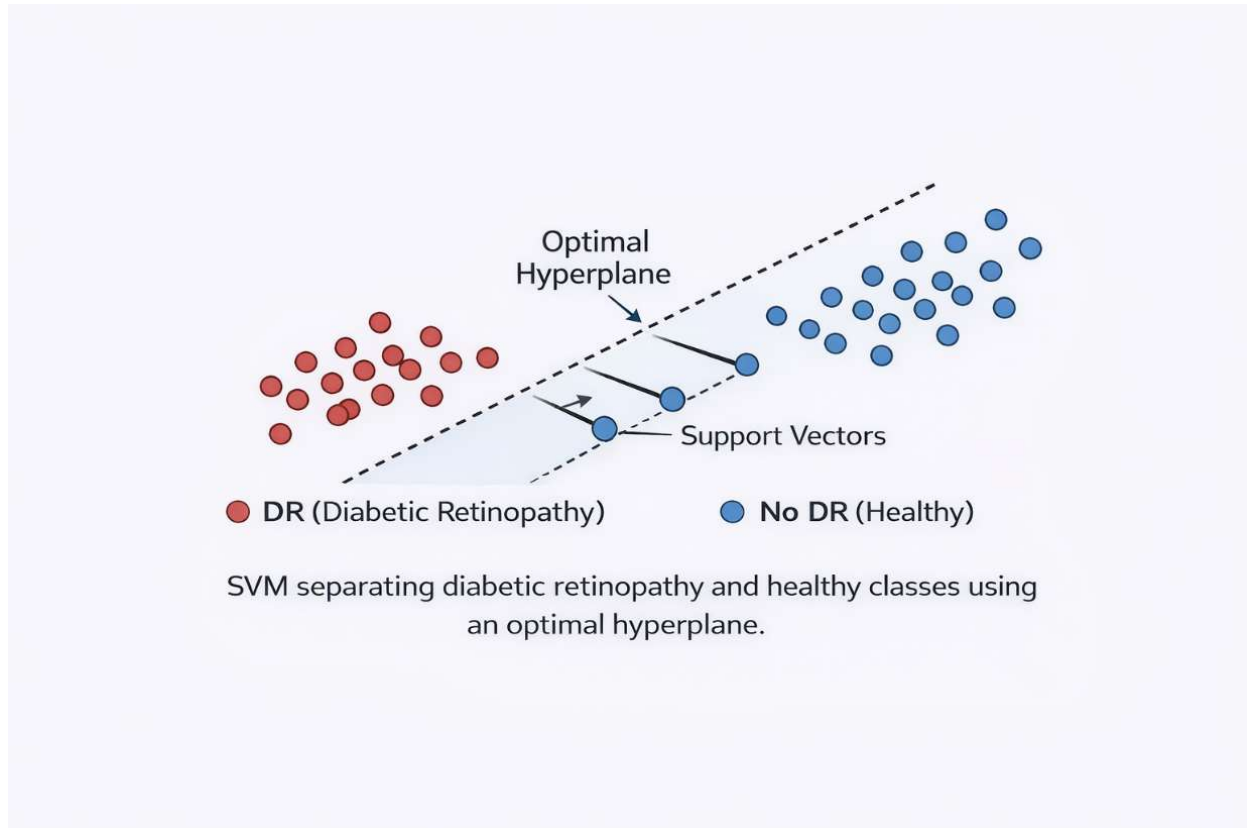


Figure 2: Support Vector Machine Classification of Retinal Features [6]

## 2. Random Forest

Random Forest is an ensemble of decision trees rather than a single tree model. During training, different trees are built from bootstrap samples and randomized subsets of the available predictors. Their outputs are then combined by voting to produce the final class. This design is useful for the present problem because it can capture nonlinear relationships, is relatively tolerant of correlated variables, and supports class weighting for imbalanced data [7]. These characteristics make it suitable for the high-dimensional fused retinal embeddings used in this thesis. Figure 3 illustrates the voting process.

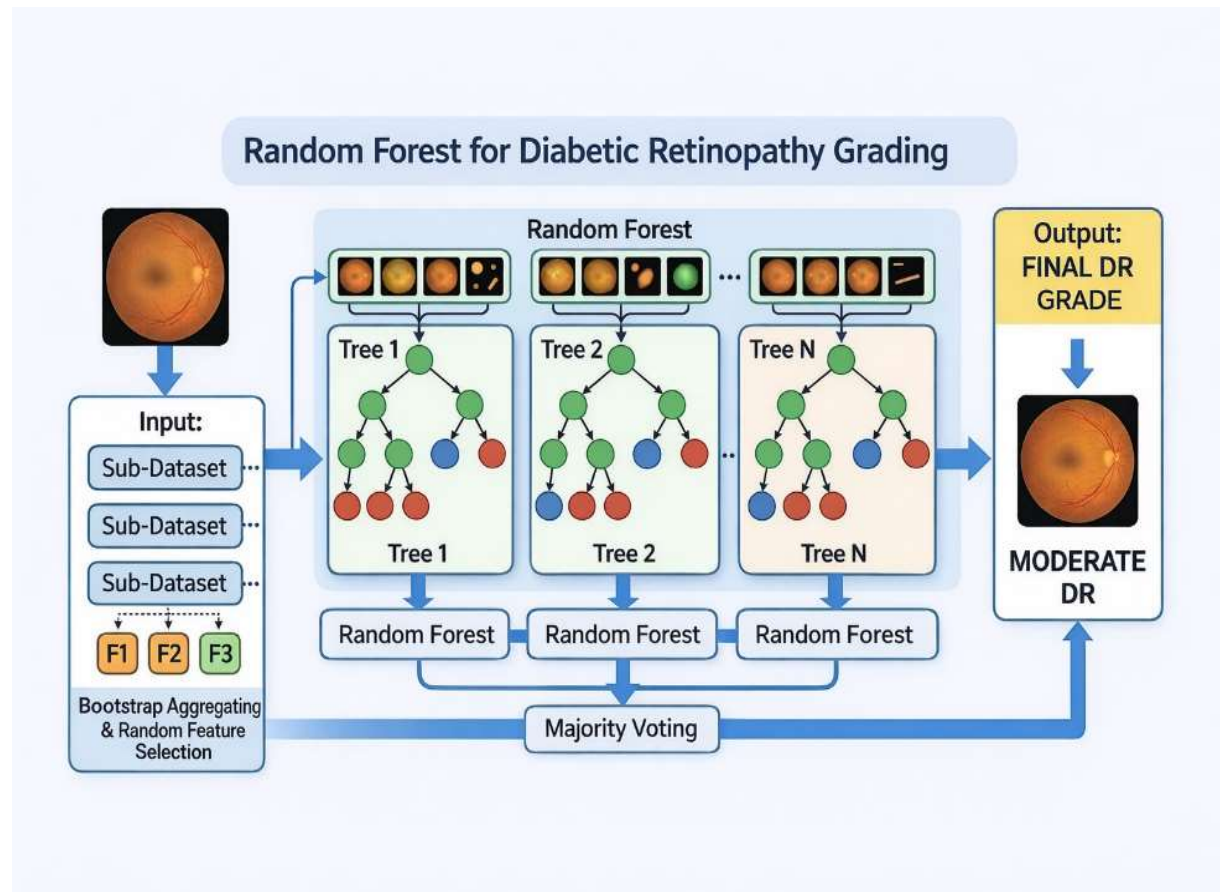


Figure 3: Random Forest Classification for Diabetic Retinopathy Grading [7]

### 3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors classifies a new sample by looking at the labels of nearby samples in feature space. The method is straightforward, but its result depends heavily on how distance is measured, how features are scaled, the selected value of  $k$ , and the dimensionality of the representation [8]. Figure 4 illustrates the neighborhood-voting idea.

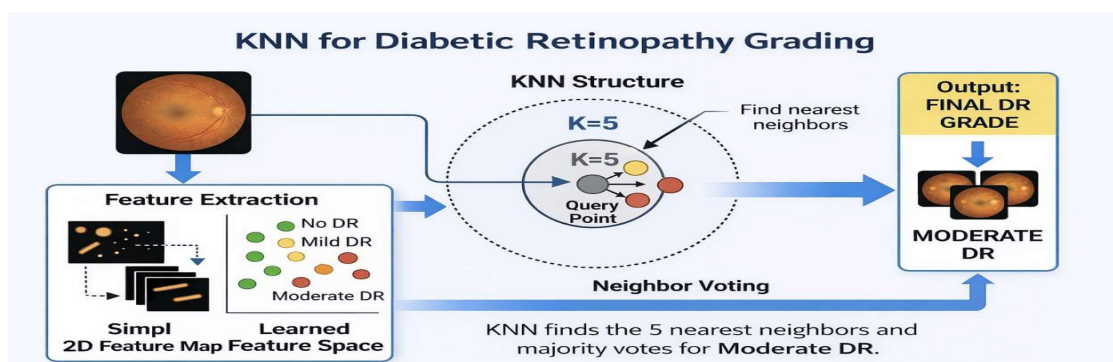


Figure 4: K-Nearest Neighbors for Diabetic Retinopathy Grading [8]

#### 2.2.4 Diabetic Retinopathy Detection Using Deep Learning

Deep networks learn visual representations directly from retinal images in a hierarchy. Earlier layers tend to respond to basic structures such as edges, color changes, and texture, while deeper layers can represent vessels, lesions, and more complex disease patterns. Their performance nevertheless depends on image quality, class balance, augmentation, model capacity, and the diversity of training examples [9]. ConvNeXt-Tiny was selected for the convolutional branch because it provides a modern CNN architecture while preserving the spatial biases that are useful for local retinal structure.

## 1. ConvNeXt-Tiny (Pre-trained)

ConvNeXt-Tiny updates traditional CNN architecture by using larger kernels, stage-wise hierarchical features, as well as different normalization and optimization techniques. Such features allow the extraction of local lesions' textures, vessels' structure, and more abstract patterns associated with the presence of diseases from retinal images. In the proposed approach, the finetuned model yields a 768-dimensional feature embedding after global averaging; the classifier is discarded, while the feature embedding is utilized for fusion [10]. Figure 5 presents the illustration of this step conceptually.

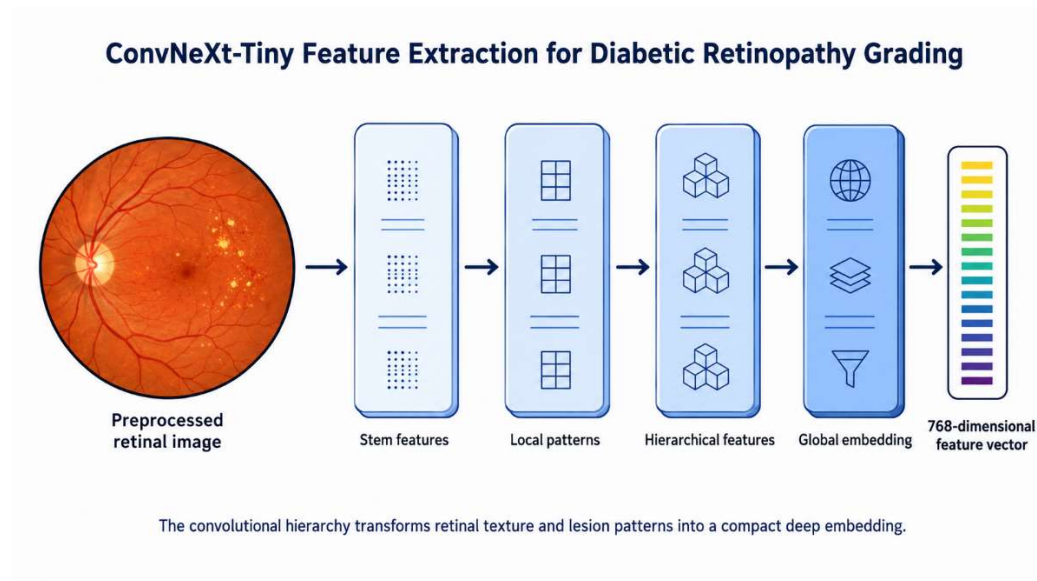


Figure 5: ConvNeXt-Tiny Feature Extraction for Diabetic Retinopathy Grading [10]

### 2.2.5 Diabetic Retinopathy Detection Using Transformer Models

Transformer-based vision models represent an image as tokens or local regions and use self-attention to relate information across space. In retinal analysis, this allows evidence from separated areas of the fundus to influence the representation, complementing convolutional models whose operations are more strongly local. These models can nevertheless be data- and compute-intensive and require

appropriate regularization [11], [12]. Figure 6 summarizes the transformer encoders considered in the background review.

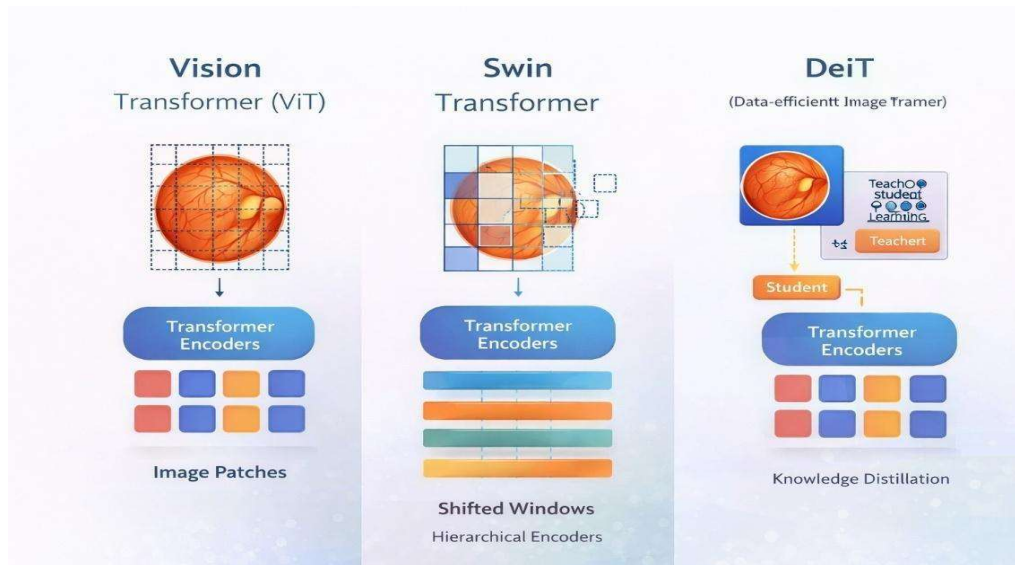


Figure 6: Transformer-Based Retinal Image Encoding [11] [12]

## 1- Vision Transformer (ViT)

Vision Transformer (ViT) is one of the earliest transformer architectures for classification of images. While CNN networks use convolutional filters for image analysis, Vision Transformer splits an image into different patches and encodes each patch as a vector referred to as patch embedding with positional encoding that encodes the spatial relations within patches. Such embeddings go through the layers of transformers' encoder where the relationships between various regions within the image are established through the attention mechanism. One of the key features of Vision Transformer is the possibility of capturing dependencies between all pixels of an entire image. Thus, Vision Transformers could be used effectively for medical imaging tasks when some important pathological alterations in organs spread throughout different locations of an image. For instance, such alterations are relevant

for the detection of diabetic retinopathy because one should analyze several regions of the retina. However, in order for ViT to function properly, one should provide more data for training and consume more computational power [13]. Figure 7 illustrates patch-based transformer encoding applied to retinal image analysis.

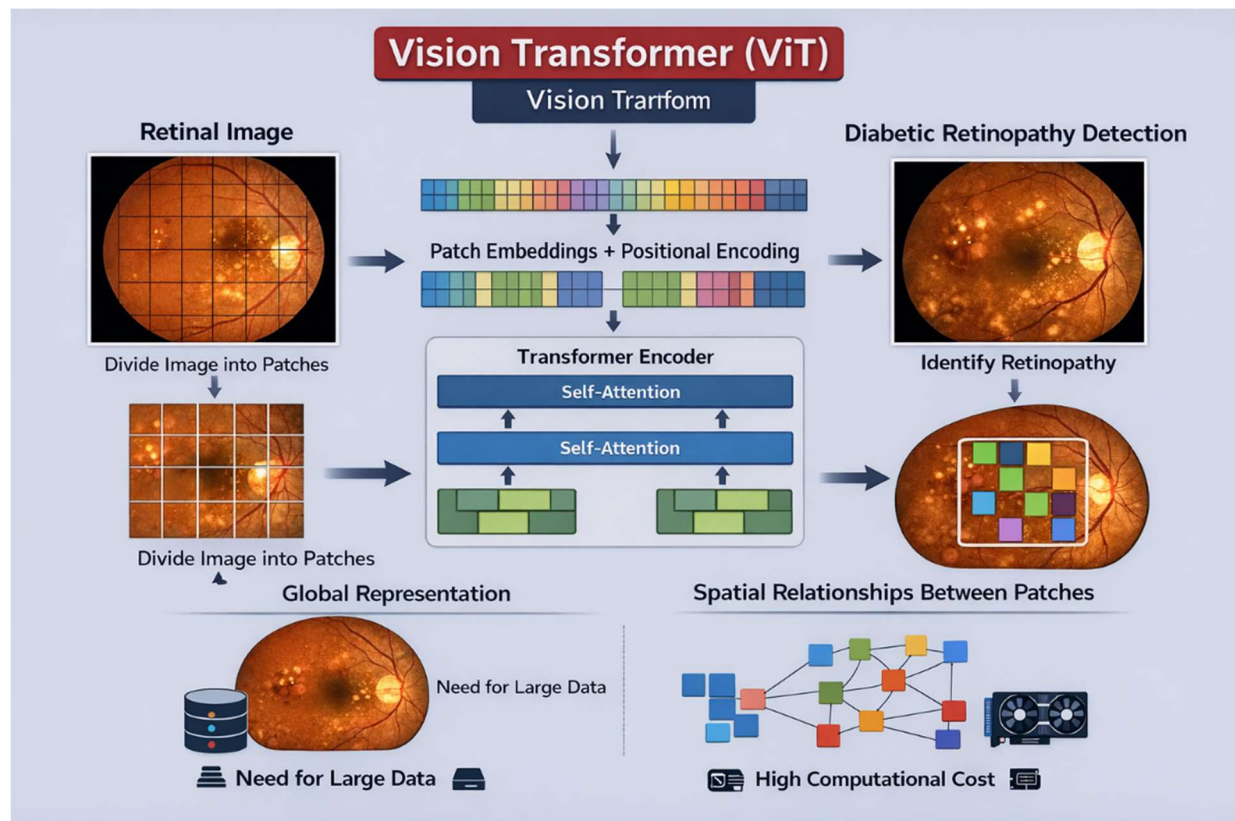


Figure 7: Vision Transformer Encoding for Diabetic Retinopathy Analysis [11]

## 2- Swin Transformer

Swin Transformer uses self-attention over local windows along with window shifting at each block. In other words, this architecture helps reduce computational cost by allowing information exchange among neighbors and different spatial levels. Therefore, Swin Transformer is appropriate to analyze retinas due to the need for fine lesions and wider patterns of lesions. The shifted window mechanism for retinopathy grading is shown in Figure 8.

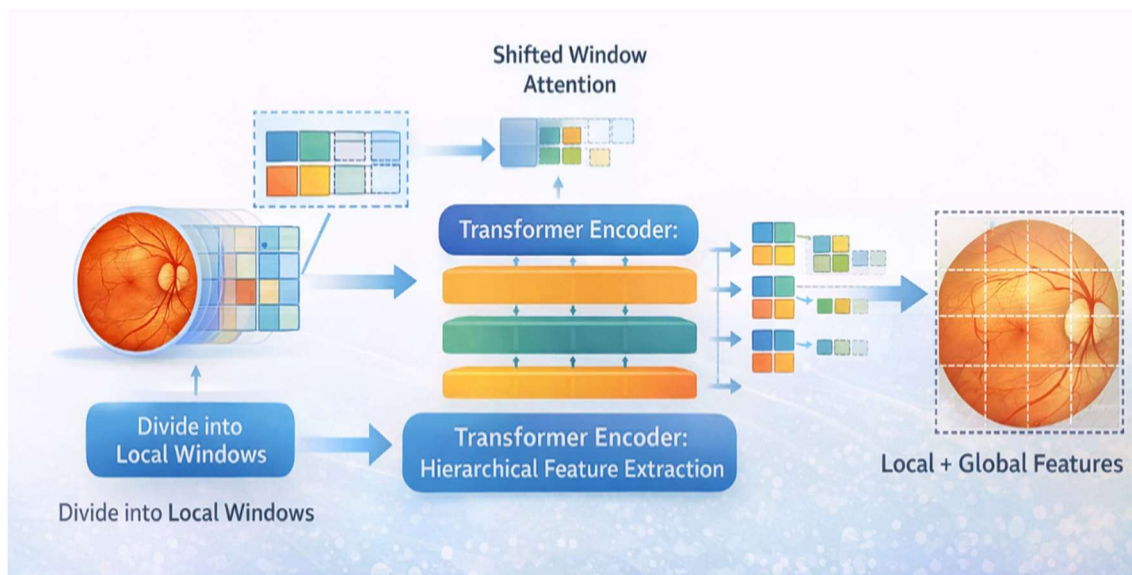


Figure 8: Swin Transformer Encoding for Diabetic Retinopathy Grading [12]

## 2.2.6 Deep Feature Fusion and Hybrid Classification

Feature-level fusion merges the features obtained using different architectures

prior to classification. In the case of retinal image analysis, ConvNeXt-Tiny may focus on the local features while Swin-V2-Tiny can capture context. Direct concatenation preserves both 768-dimensional embeddings and forms a 1,536-dimensional vector for the downstream Random Forest classifier. Regularization is required because fusion increases dimensionality and may introduce redundant variables [7], [10], [14], [15]. The adopted process is summarized in Figure 9.

Direct concatenation is used instead of averaging because the ConvNeXt-Tiny and Swin-V2-Tiny embeddings are not assumed to occupy identical feature spaces. Concatenation retains all architecture-specific variables, while the regularized Random Forest can select informative nonlinear interactions from the fused representation [7].

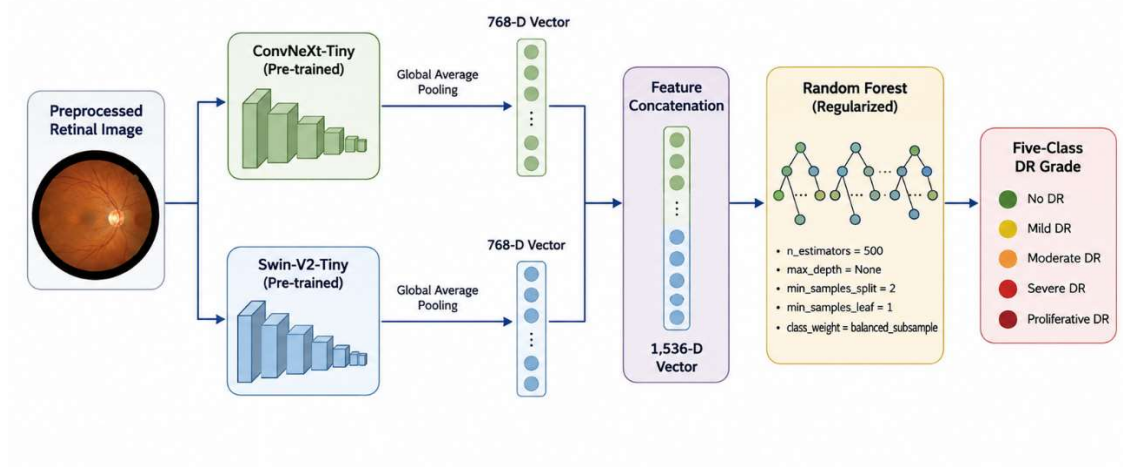


Figure 9: ConvNeXt-Swin Feature Fusion and Random Forest Classification [7] [10] [14] [15]

## 2.2.7 Performance Metrics

A single accuracy value can hide important weaknesses in a five-class diabetic retinopathy model, especially when common classes contribute far more samples than rare severe grades. For that reason, the evaluation in this study reports accuracy together with class-wise precision, recall, F1-score, specificity, confusion-matrix results, and one-vs-rest ROC-AUC. Macro averages give each grade the same weight, whereas weighted averages reflect the actual class frequencies. Considering both makes it possible to see whether a strong overall result is also supported by reasonable performance on the minority classes.

### ● Accuracy

Accuracy expresses the fraction of all evaluated images that receive the correct predicted label.

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \quad (1)$$

- **Precision**

Precision asks, for a given class, what proportion of the images predicted as that class truly belong to it.

$$Precision = \frac{TP}{TP+F} \quad (2)$$

- **Recall (Sensitivity)**

Recall (sensitivity) measures how many of the actual examples of a class are recovered by the classifier.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- **F1-Score**

The F1-score summarizes the balance between precision and recall by taking their harmonic mean; it falls when either quantity is poor.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Reca} \quad (4)$$

- **Specificity**

Specificity measures the model's ability to correctly reject images that do not belong to the class being considered.

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

- **Confusion Matrix**

A confusion matrix places the true and predicted classes in one table, making both correct decisions and recurring class-to-class errors visible.

For each class treated in a one-vs-rest manner, the matrix can be used to obtain TP, TN, FP, and FN counts.

- **ROC Curve and AUC**

The ROC curve shows how sensitivity changes relative to the false-positive rate as the classification threshold is varied. The area under that curve (AUC) summarizes how well the model ranks positive examples ahead of negative ones; larger values indicate better separation.

- **AUC = 1** indicates perfect classification
- **AUC = 0.5** indicates random performance

For this five-class task, no metric is interpreted alone. A model can look strong because it performs well on No DR or Moderate while still missing many Mild, Severe, or Proliferative DR cases. Class-specific values and macro averages are therefore examined alongside overall measures to judge whether performance is distributed across the full severity range.

## 2.3 Literature Review

Grzybowski and Jin [13] reviewed 13 CE-approved systems used in Europe for automated or semi-automated diabetic retinopathy detection. Most relied on deep learning with fundus photographs, but the systems were validated using different populations, imaging modalities, and study designs. Because of that heterogeneity, the authors did not combine the results into a single pooled estimate or directly rank the systems. Reporting for sight-threatening DR was also less consistent, with controlled evaluations averaging roughly 80% performance where such results were available. The review is useful for understanding regulatory and implementation issues, although its focus is commercial screening systems rather than an experimental five-class grading model.

Akhtar et al. [16] evaluated RSG-Net on Messidor-1 for both a four-stage grading problem and a binary classification task. Their pipeline included lesion-focused cropping, denoising, histogram equalization, resizing, augmentation, and an optimized convolutional model. They reported test accuracies of 99.36% for four-stage grading and 99.37% for the binary task. These results are high, but the near-perfect training behavior and use of a single dataset make external generalization difficult to judge.

Guefrachi et al. [17] compared a staged transfer-learning strategy across several pretrained CNNs, including InceptionResNetV2, VGG16, VGG19, DenseNet121, MobileNetV2, and EfficientNetV2L. Data augmentation was part of training, and the best configuration reached about 96.61% accuracy for diabetic retinopathy detection and grading. The study supports the value of transfer learning, although the relatively limited dataset size and absence of broad external testing restrict conclusions about performance in other clinical settings.

Youldash et al. [18] evaluated six pretrained CNN families on APTOS and on retinal images collected from a Saudi clinical center. On APTOS, RegNetX080 achieved 98.6% accuracy for binary classification, while EfficientNetB3 achieved 85.1% for multiclass grading. In the external clinical experiment, the best binary and multiclass accuracies were 98.2% and 84.4%, respectively. The study provides useful evidence that binary detection is substantially easier than five-class grading and highlights the continuing effects of class imbalance and domain variation.

Shamrat et al. [19] introduced DRNet13, a custom convolutional network for five-class diabetic retinopathy grading, and compared it with fifteen pretrained CNN architectures. The APTOS images were enhanced using median filtering and gamma correction, and augmentation increased the dataset from 3,662 to 7,500 images. DRNet13 achieved 97% accuracy and was reported to be computationally efficient relative to the evaluated backbones. However, the results were obtained from a single public dataset, and the absence of independent clinical validation limits conclusions regarding deployment robustness.

Jabbar et al. [14] combined lesion-oriented deep features from GoogleNet and ResNet, applied adaptive particle swarm optimization, and then compared several conventional classifiers such as Random Forest, SVM, Decision Tree, and Linear Regression. The best reported benchmark accuracy was 94%. Their work demonstrates that deep representations can be paired with classical classifiers, but the paper provides limited evidence about external validation and full reproducibility of the model-selection process.

Fayyaz et al. [20] extracted representations from AlexNet and ResNet101, fused them, reduced the feature set with an ant-colony procedure, and compared several SVM kernels. With 750 selected features, the cubic SVM reached 93.0% accuracy, 93.0% precision, 93.5% recall, and a 93.2% F1-score. This study illustrates the

benefit of combining deep features with a conventional classifier, while the multi-stage feature-selection process and restricted evaluation setting may make independent reproduction and generalization more difficult.

Alwakid et al. [21] tested DenseNet121 on APTOS 2019 under several image-enhancement settings based on histogram equalization, CLAHE, and ESRGAN. Their CLAHE-plus-ESRGAN configuration produced 98.36% five-class test accuracy, and the reported top-2 and top-3 accuracies were 100%. The work shows that strong enhancement can improve classification performance, but it also increases computational processing and was evaluated only on one public dataset.

that contrast enhancement can be used to boost retinal image classification; however, this method requires heavy computational enhancement and is only validated through one public data set.

Lin and Wu [22] modified ResNet50 and paired it with a standardized preprocessing routine and an adaptive learning-rate schedule. Their model achieved 83.95% training accuracy and 74.32% test accuracy, outperforming the conventional CNN baselines reported in the same study while showing less training instability. The remaining difference between training and test performance suggests that generalization was still limited.

Mohanty et al. [23] made a comparison of a hybrid VGG16-XGBoost approach with DenseNet121 on the APTOS 2019 dataset after preprocessing and class balancing. While the hybrid architecture scored 79.50% accuracy, DenseNet121 scored 97.30%, indicating that the hybrid architecture does not necessarily perform better than an end-to-end architecture. The importance of the chosen feature extractor and fusion method is highlighted in this paper, but the study lacked independent external validation.

Patil et al. [24] have used the ResNet50 architecture on the APTOS dataset and performed experiments on the EyePACS dataset by using image pre-processing, augmentation, and test time augmentation techniques to achieve good performance. They obtained an accuracy of 97.87% and a quadratic weighted kappa value of 0.985. The cross-dataset methodology is a strong point because it tests for robustness outside the training dataset; yet, it requires extensive augmentation and specific qualities of the chosen external subset.

Ali and Dawood [25] optimized ResNet50 to perform diabetic retinopathy grading with APTOS 2019 fundus images. Preprocessing, transfer learning, checkpoint selection, and multi-class evaluation were their pipeline stages. The obtained accuracy was 98.3%, the precision reached 98.4%, the recall amounted to 98.4%, and the F1-score made 98.5%. Despite the great accuracy inside the dataset, the lack of external validation makes it hard to estimate the generalization capabilities of the model on other images.

Das et al. [26] proposed the DRFEC benchmark and tested 26 architectures of pretrained CNNs in feature extraction and classification of the EyePACS dataset. EfficientNetB4 achieved the best validation accuracy of 79.11% against the training accuracy of 99.37%, while other larger models showed significant overfitting. The architectural comparison is helpful, yet wide gaps between training and validation indicate that even high model capacity and good training accuracy are not enough for the generalization.

Jiwani et al. [27] utilized a convolutional neural network for classifying diabetic retinopathy using the IDRiD data set. Based on the later reporting, the accuracy of this method is about 77.6%, which shows relatively decent results in terms of a relatively small data set of retinas. This article provides a nice starting point

for CNN-based grading; however, poor reporting of the methodology makes the reproducing impossible.

## 2.4 Comparative Analysis of Related Work

Taken together, the fourteen studies reveal two recurring challenges. First, retinal datasets vary in acquisition quality and usually contain an uneven number of examples across disease grades, so preprocessing and augmentation are common design choices. Second, a high result on one dataset should not automatically be interpreted as evidence of equal performance on another source of images, particularly for multiclass grading. The reviewed work also shows a wide range of modeling strategies, from end-to-end CNNs to transformer models, feature fusion, and classical classifiers, which supports evaluating a hybrid representation under a clearly separated validation and test protocol.

One area that remains less extensively examined is the combination of a recent convolutional backbone with a hierarchical vision transformer at the feature level, followed by a deliberately regularized ensemble classifier. The present study addresses that gap by concatenating ConvNeXt-Tiny and Swin-V2-Tiny embeddings and evaluating a Random Forest under separate stratified training, validation, and test partitions.

*Table 1: Comparative study of diabetic retinopathy detection methods*

| Authors                 | Year | Method / Model                            | Dataset                         | Results                          | Limitations                                |
|-------------------------|------|-------------------------------------------|---------------------------------|----------------------------------|--------------------------------------------|
| Grzybowski and Jin [13] | 2026 | Review of 13 CE-certified AI systems      | EU validation studies           | No pooled estimate; STDR ~80%    | Heterogeneous validation and reporting     |
| Akhtar et al. [16]      | 2025 | RSG-Net                                   | Messidor-1                      | 99.36% four-stage; 99.37% binary | Single dataset; overfitting risk           |
| Guefrachi et al. [17]   | 2025 | Multistage transfer learning              | Kaggle fundus images            | 96.61% accuracy                  | Limited data; no broad external validation |
| Youldash et al. [18]    | 2024 | EfficientNet and RegNet transfer learning | APTOS and Saudi clinical images | 98.6% binary; 85.1% multiclass   | Lower multiclass performance               |
| Shamrat et al. [19]     | 2024 | DRNet13 vs. 15 pretrained CNNs            | APTOS 2019                      | 97% accuracy                     | Single public dataset                      |
| Jabbar et al. [14]      | 2024 | GoogleNet-ResNet + APSO + ML              | Fundus dataset                  | 94% accuracy                     | Limited external validation                |
| Fayyaz et al. [20]      | 2023 | AlexNet-ResNet101 fusion + SVM            | EyePACS                         | 93.0% accuracy; 93.2% F1-score   | Complex feature-selection pipeline         |
| Alwakid et al. [21]     | 2023 | DenseNet121 + CLAHE/ESRGAN                | APTOS 2019                      | 98.36% five-class accuracy       | Heavy enhancement; single dataset          |
| Lin and Wu [22]         | 2023 | Revised ResNet50                          | Kaggle DR images                | 83.95% train; 74.32% test        | Remaining train-test gap                   |
| Mohanty et al. [23]     | 2023 | VGG16-XGBoost vs. DenseNet121             | APTOS 2019                      | 79.50% vs. 97.30%                | No external validation                     |
| Patil et al. [24]       | 2023 | ResNet50 + augmentation/TTA               | APTOS to EyePACS                | 97.87%; QWK 0.985                | Extensive augmentation                     |
| Ali and Dawood [25]     | 2023 | Fine-tuned ResNet50                       | APTOS 2019                      | 98.3% accuracy; 98.5% F1-score   | No external validation                     |
| Das et al. [26]         | 2023 | DRFEC: 26 pretrained CNNs                 | EyePACS                         | Best validation : 79.11%         | Overfitting and high train-validation gap  |
| Jiwani et al. [27]      | 2022 | CNN classifier                            | IDRiD                           | ~77.6% accuracy                  | Limited design detail                      |

## 2.5 Chapter Summary

This chapter presented the medical and computational foundations required for automated diabetic retinopathy grading. It described the five clinical severity classes and reviewed the roles of preprocessing, augmentation, transfer learning, machine-learning classifiers, convolutional networks, and vision transformers in retinal-image analysis. It also explained the evaluation measures needed for imbalanced multiclass classification. Fourteen related studies were examined to compare their datasets, methods, reported results, and limitations. The review indicates that strong within-dataset performance does not consistently translate into cross-dataset generalization and that comparatively little work has combined complementary CNN and transformer embeddings with a regularized machine-learning classifier.

## Chapter 3: Proposed Approach (Methodology)

### 3.1 Introduction

Chapter 3 details how the proposed five-class system was implemented and evaluated. It covers the APTOS 2019 data, fixed preprocessing, stratified splitting, augmentation limited to training, the two deep feature extractors, concatenation of their embeddings, and the procedure used to choose the Random Forest classifier.

### 3.2 Proposed Framework

The proposed architecture integrates convolutional and transformer-based visual representations. ConvNeXt-Tiny supplies a strong local inductive bias for retinal texture, vessel structure, exudates, hemorrhages, and other lesion-related patterns. Swin-V2-Tiny complements these features through hierarchical shifted-window self-attention, which captures broader spatial relationships and contextual dependencies across the fundus image. The penultimate 768-dimensional output of each backbone is retained, producing two complementary embeddings for every image.

Direct concatenation forms a 1,536-dimensional hybrid feature vector without averaging or discarding architecture-specific information. A Random Forest then performs the final five-class prediction. Separating deep feature learning from the final machine-learning classifier allows the optimized embeddings to be stored and reused, while validation-based model selection limits direct dependence on the held-out testing subset. The complete architecture of the proposed framework is summarized in Figure 10.

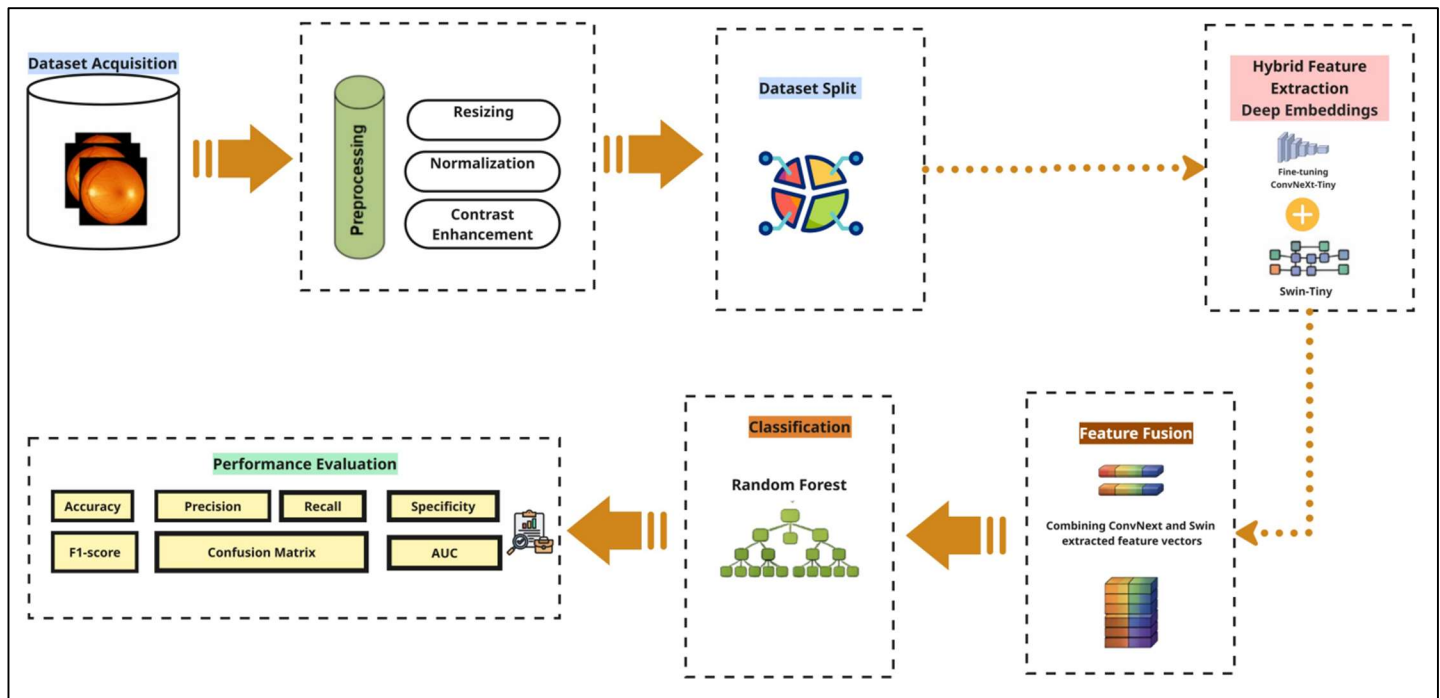
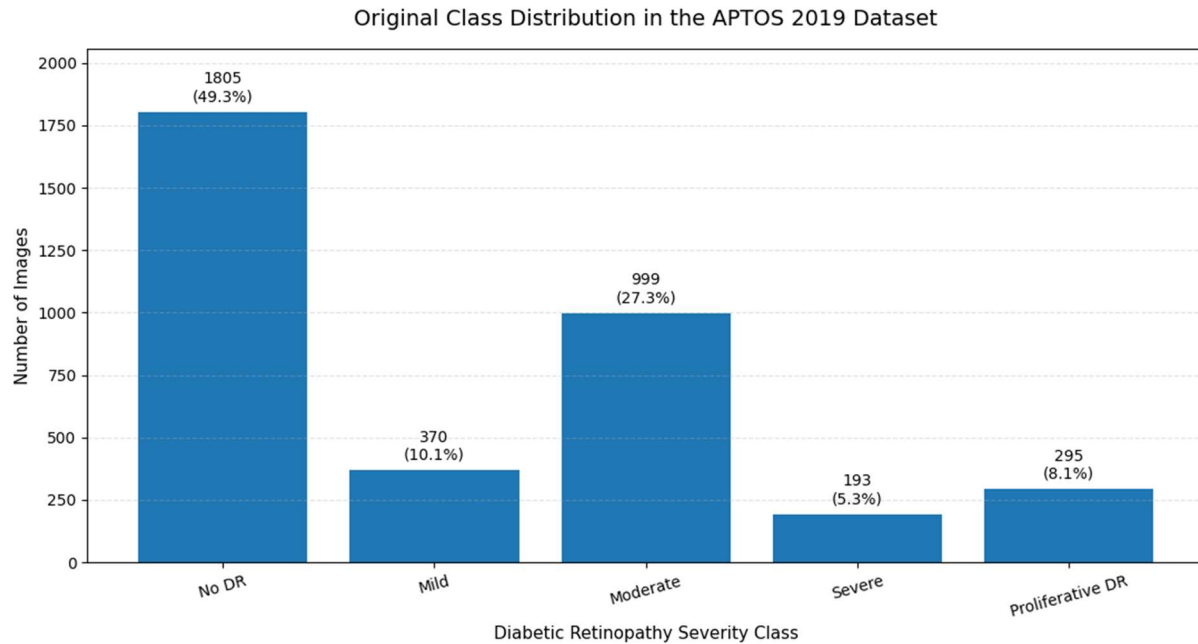


Figure 10: Proposed Hybrid Framework for Diabetic Retinopathy Classification

### 3.3 Dataset Description

This study uses the APTOS 2019 Blindness Detection dataset, containing 3,662 color fundus photographs distributed across five diagnostic labels: No DR, Mild, Moderate, Severe, and Proliferative DR. Before any model training, the image files were reconciled against the label metadata. Checks covered missing images, repeated identifiers, labels outside the expected range, and mismatches between filenames and the accompanying records. The organized dataset contained no missing image files.

Class frequencies are strongly uneven: No DR dominates the dataset, while Severe and Proliferative DR are comparatively scarce. The experimental design therefore preserves class proportions during splitting, uses class-sensitive training choices, and reports both macro and frequency-weighted metrics. Figure 11 shows the original counts.



*Figure 11: Original Class Distribution in the APTOS 2019 Dataset*

Figure 12 shows one original image selected from each severity class before preprocessing. The examples are included to make the raw variation in the dataset visible, including differences in image dimensions, illumination, dark borders, retinal field of view, and overall appearance.



*Figure 12: Representative Retinal Images Before Preprocessing*

### 3.4 Data Preprocessing

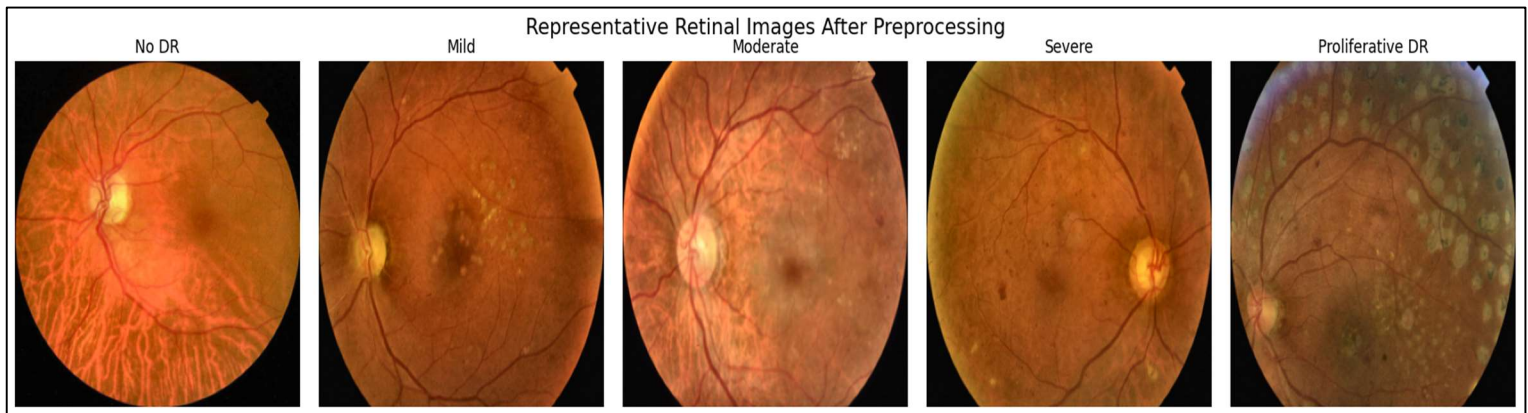
The same deterministic preprocessing pipeline was applied to the training, validation, and test subsets. Random transformations were introduced only after this step and only for training. The objective of deterministic preprocessing was to reduce variation unrelated to disease severity while preserving retinal structures needed for grading. Dark peripheral areas were removed using a foreground mask derived from a grayscale representation. The cropped image was converted to LAB color space, and CLAHE was applied to the lightness channel to improve local contrast without excessively amplifying noise.

After contrast enhancement, every image was resized to 300 x 300 pixels using area interpolation, converted to an RGB float32 PyTorch tensor, and normalized with the ImageNet channel means and standard deviations expected by the pretrained backbones. Table 2 lists the exact values used in the implementation so that the preprocessing procedure can be reproduced.

Table 2: Deterministic Retinal-Image Preprocessing Parameters

| Parameter                        | Value                  |
|----------------------------------|------------------------|
| Foreground threshold             | 10                     |
| Minimum crop size                | 50 pixels              |
| CLAHE clip limit                 | 1.5                    |
| CLAHE tile grid                  | 8 x 8                  |
| Image size                       | 300 x 300 pixels       |
| Interpolation                    | INTER_AREA             |
| Tensor type                      | float32 PyTorch tensor |
| Normalization mean               | [0.485, 0.456, 0.406]  |
| Normalization standard deviation | [0.229, 0.224, 0.225]  |

Figure 13 gives some representative images after the deterministic pre-processing stage. Unlike the original sample images, the new pre-processed images are standardized in size, minimize border areas, increase local contrast, while still capturing the main features of the retinal anatomy needed for feature extraction.



*Figure 13: Representative Retinal Images After Preprocessing*

## 3.5 Dataset Splitting and Training-Only Augmentation

### 3.5.1 Stratified Dataset Partitioning

A fixed 70:15:15 stratified partition was applied to the full dataset, resulting in 2,563 training, 549 validation, and 550 test images. Stratification preserved the relative frequency of each diagnostic grade across the three subsets. Validation data were used for checkpoint and hyperparameter decisions, while the test subset remained untouched until model selection was complete. Class-level counts and their effects are examined in Chapter 4.

### 3.5.2 Training-Only Data Augmentation

Only the training subset received stochastic augmentation. For ConvNeXt-Tiny, the pipeline included horizontal flips, small rotations, mild affine shifts and scaling, brightness and contrast perturbations, saturation variation, and random erasing. Swin-V2-Tiny was trained with a lighter set of transformations after the stronger initial setting was found to limit learning. Validation and test images underwent deterministic preprocessing and normalization only, so their reported scores were not influenced by random transformations.

## 3.6 Deep Feature Extraction

Because APTOS 2019 is far smaller than the corpora typically used to learn modern visual backbones from scratch, both networks were initialized from ImageNet-1K weights and then fine-tuned on the five diabetic retinopathy classes. Once the best validation checkpoints were selected, the classification heads were bypassed and the penultimate-layer representations were generated for all training, validation, and testing images.

### 3.6.1 ConvNeXt-Tiny

ConvNeXt-Tiny was fine-tuned on the training subset and used as the convolutional feature extractor. The selected checkpoint was determined from validation performance. After fine-tuning, the classification head was removed and a 768-dimensional embedding was extracted for each image. Table 3 lists the final implemented parameters.

Table 3: ConvNeXt-Tiny Final Parameters

| Parameter             | Value                              |
|-----------------------|------------------------------------|
| Pretrained weights    | ImageNet-1K V1                     |
| Input size            | 300 x 300 pixels                   |
| Batch size            | 32                                 |
| Output classes        | 5                                  |
| Embedding dimension   | 768                                |
| Classifier head       | Dropout (0.45) + Linear (768 to 5) |
| Loss function         | Class-weighted cross-entropy       |
| Label smoothing       | 0.06                               |
| Optimizer             | AdamW                              |
| Best checkpoint epoch | 15                                 |

### 3.6.2 Swin-V2-Tiny

Swin-V2-Tiny was fine-tuned as the complementary transformer feature extractor. Its shifted-window hierarchy captured contextual retinal information that differs from convolutional representations. The classification head was removed after checkpoint selection, producing a 768-dimensional embedding per image. Table 4 summarizes the final settings.

Table 4: Swin-V2-Tiny Final Parameters

| Parameter              | Value                              |
|------------------------|------------------------------------|
| Pretrained weights     | ImageNet-1K V1                     |
| Input size             | 300 x 300 pixels                   |
| Batch size             | 32                                 |
| Output classes         | 5                                  |
| Embedding dimension    | 768                                |
| Classifier head        | Dropout (0.20) + Linear (768 to 5) |
| Loss function          | Class-weighted cross-entropy       |
| Label smoothing        | 0.02                               |
| Backbone learning rate | $1 \times 10^{-5}$                 |
| Head learning rate     | $2 \times 10^{-4}$                 |
| Optimizer              | AdamW                              |
| Best checkpoint epoch  | 10 (refinement stage)              |

## 3.7 Feature Fusion

The two deep representations were fused by concatenating their feature dimensions. Each image therefore contributed 768 ConvNeXt values and 768 Swin values, yielding a single 1,536-element vector. The features were not averaged and no dimensionality-reduction step was introduced before Random Forest training. This

choice keeps the information produced by each backbone separate within the fused vector and leaves the classifier to determine which variables and nonlinear combinations are useful.

After extraction, the fused matrices for training, validation, and testing were stored with their labels as compressed NumPy files. Persisting these embeddings avoids repeating the expensive backbone inference stage during later machine-learning experiments and makes it possible to reconstruct the final classifier after a session reset using exactly the same input representation.

### 3.8 Random Forest Classification

Random Forest was trained on the 1,536-dimensional fused vectors. Candidate hyperparameter settings were explored by randomized search with stratified cross-validation on the training portion, and the validation subset was used to verify the preferred regularized configuration. Test data were excluded from this selection process. The parameters used for the final held-out evaluation are listed in Table 5.

*Table 5: Selected Random Forest Parameters*

| <i>Parameter</i>             | <i>Value</i>              |
|------------------------------|---------------------------|
| <i>Number of trees</i>       | <i>700</i>                |
| <i>Bootstrap</i>             | <i>True</i>               |
| <i>Class weighting</i>       | <i>balanced subsample</i> |
| <i>Criterion</i>             | <i>entropy</i>            |
| <i>Maximum depth</i>         | <i>3</i>                  |
| <i>Minimum samples split</i> | <i>80</i>                 |
| <i>Minimum samples leaf</i>  | <i>20</i>                 |
| <i>Maximum features</i>      | <i>sqrt</i>               |
| <i>Maximum samples</i>       | <i>0.25</i>               |
| <i>Random seed</i>           | <i>42</i>                 |
| <i>Cross-validation</i>      | <i>Stratified 5-fold</i>  |

### 3.9 Chapter Summary

Chapter 3 defined the full experimental pipeline, from APTOS 2019 preprocessing and the stratified 70:15:15 partition to dual-backbone embedding extraction and feature fusion. ConvNeXt-Tiny and Swin-V2-Tiny each produced 768 features per image; concatenation created the 1,536-dimensional input used by Random Forest. The classifier configuration was fixed through training and validation procedures before the test subset was evaluated.

## Chapter 4: Experimental Results and Discussion

### 4.1 Introduction

This chapter reports the held-out evaluation of the hybrid system. The main analysis concerns the Random Forest trained on the concatenated ConvNeXt-Tiny and Swin-V2-Tiny embeddings. Performance is examined through overall and class-sensitive metrics, confusion patterns, one-vs-rest ROC-AUC, and a contextual comparison with selected previous studies. The final sections interpret what the results suggest about feature fusion, regularization, class imbalance, and generalization.

### 4.2 Experimental Setup

All experiments were carried out in Google Colab with Python and the PyTorch, Torchvision, OpenCV, and scikit-learn libraries. GPU resources were used for backbone fine-tuning and embedding extraction, whereas the Random Forest search and final scoring were handled in scikit-learn. Random seeds were fixed in components that support deterministic control. For the final analysis, the selected backbone checkpoints and previously stored fused feature matrices were reloaded so that all reported metrics were produced from the same model state and feature representation.

The 3,662 images followed the stratified split summarized in Table 6, producing separate training, validation, and testing subsets. Random image transformations were applied exclusively to training samples. Validation accuracy determined the retained deep-learning checkpoints, while Random Forest selection considered cross-validated macro F1 together with accuracy and the size of the training-validation gap. Only after these choices were fixed was the test set used for final reporting, which includes accuracy, macro precision, macro recall, macro and weighted F1, specificity, confusion-matrix counts, and one-vs-rest ROC-AUC.

Table 6: Experimental Dataset Split

| Subset     | Number of Images | Percentage |
|------------|------------------|------------|
| Training   | 2,563            | 70%        |
| Validation | 549              | 15%        |
| Testing    | 550              | 15%        |

### 4.3 Results of the Final Hybrid Model

The final Random Forest classifier was trained on the fused 1,536-dimensional training features after cross-validated model selection. It achieved 98.95% training accuracy, 86.16% validation accuracy, and 84.91% testing accuracy. On the held-out testing subset, macro precision was 73.93%, macro recall was 68.99%, macro F1-score was 70.98%, weighted F1-score was 84.45%, macro specificity was 96.11%, macro ROC-AUC was 94.80%, and weighted ROC-AUC was 96.84%. The 14.04-percentage-point training-testing gap indicates remaining overfitting, while the close validation and testing accuracies support stable held-out generalization. Table 7 summarizes the complete set of performance measures.

Table 7: Comprehensive Performance of the Final Hybrid Framework

| Metric            | Value  |
|-------------------|--------|
| Testing accuracy  | 0.8491 |
| Macro precision   | 0.7393 |
| Macro recall      | 0.6899 |
| Macro F1-score    | 0.7098 |
| Weighted F1-score | 0.8445 |
| Macro specificity | 0.9611 |
| Macro ROC-AUC     | 0.9480 |
| Weighted ROC-AUC  | 0.9684 |

Figure 14 summarizes the final confusion matrix. Of the 550 test images, 467 fall on the diagonal and were therefore classified correctly. By class, correct predictions were 266/271 for No DR, 33/56 for Mild, 131/150 for Moderate, 14/29 for Severe, and 23/44 for Proliferative DR. The error distribution is concentrated around neighboring severity levels rather than showing frequent jumps between the mildest and most advanced categories.

The main errors occurred between neighboring grades. Mild images were frequently predicted as Moderate, and Proliferative DR images were also sometimes assigned to Moderate; Severe images showed overlap with both Moderate and Proliferative DR. Two factors help explain this pattern. The minority grades provide fewer training examples, and the visual manifestations of diabetic retinopathy evolve gradually rather than changing abruptly between adjacent labels. In addition, each fundus image is paired with a single global grade, so the training target does not explicitly identify the retinal region that contains the lesion evidence responsible for that grade.

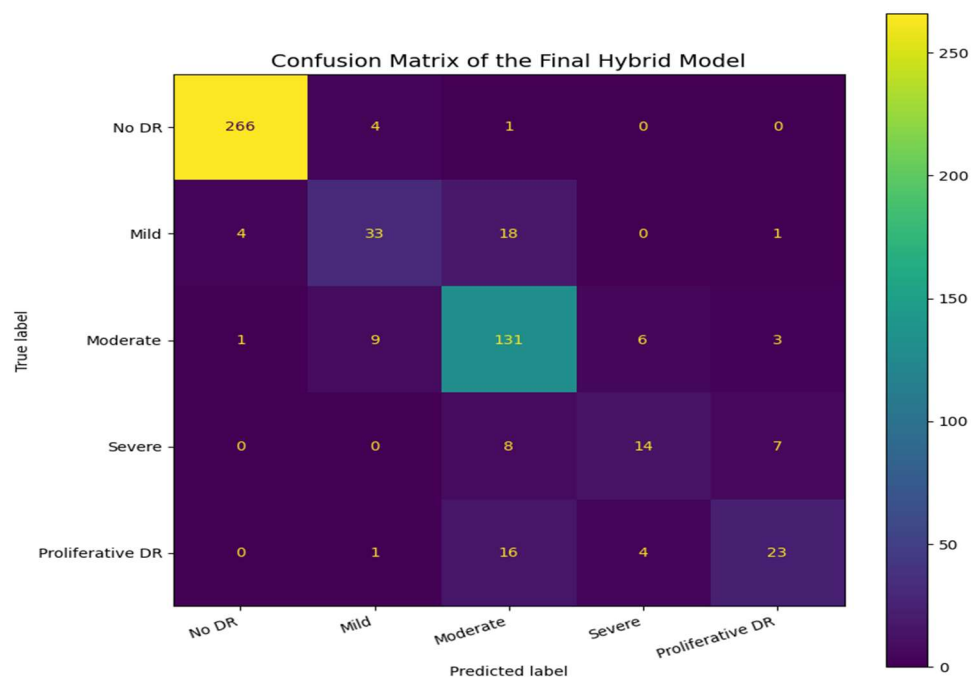


Figure 14: Confusion Matrix of the Final Hybrid Model

Figure 15 shows the one-vs-rest ROC curves derived from the Random Forest probability outputs. The AUC values were 0.9963 for No DR, 0.9271 for Mild, 0.9506 for Moderate, 0.9340 for Severe, and 0.9321 for Proliferative DR. Every curve remains clearly above the random diagonal. This indicates that the probability scores generally rank positive examples above negative examples even in classes where recall at the final decision threshold is lower.

Across the five classes, macro ROC-AUC is 0.9480, while frequency weighting increases the value to 0.9684. The weighted result is higher because the more common No DR and Moderate categories also achieve strong ranking performance. For this reason, ROC-AUC is interpreted together with recall, F1-score, specificity, and the confusion matrix rather than as a stand-alone indicator.

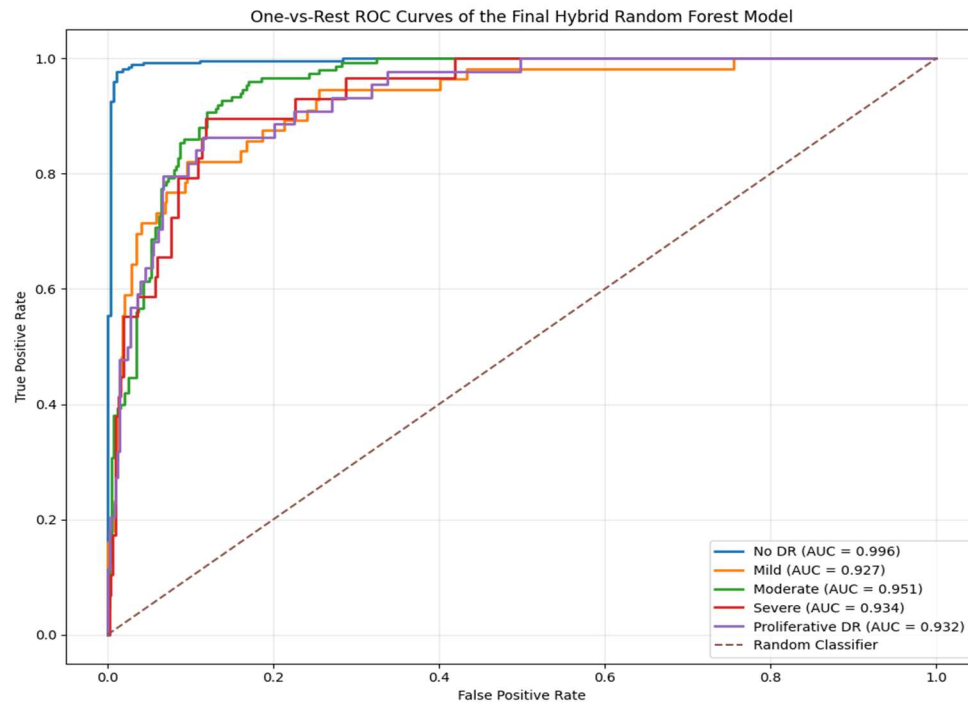


Figure 15: One-vs-Rest ROC Curves of the Final Hybrid Model

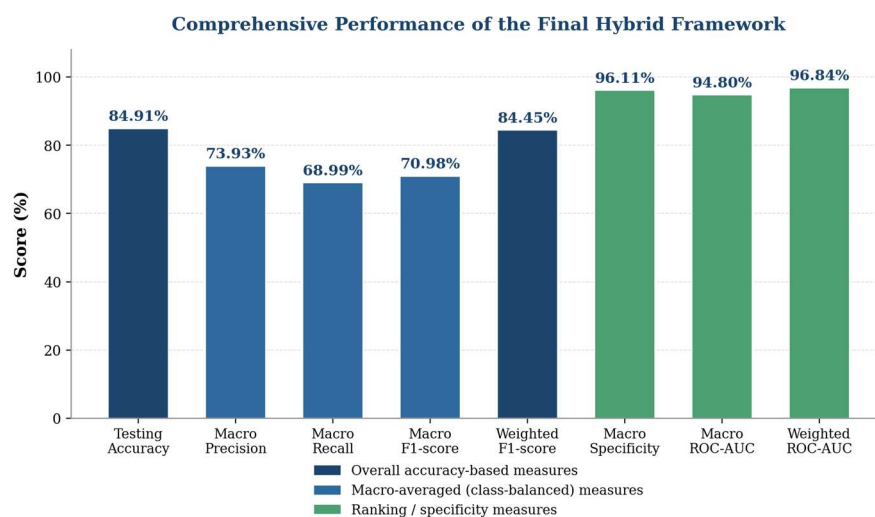


Figure 16: Bar Chart of the Comprehensive Performance Metrics of the Final Hybrid Framework

Table 8 compares the proposed model with selected studies that also used APTOS 2019 for five-class diabetic retinopathy grading. Limiting the comparison to the same dataset and number of classes makes the results more relevant, but it does not make the studies directly equivalent. Differences remain in preprocessing, data partitioning, augmentation, feature learning, and validation procedures. The table should therefore be treated as a contextual benchmark rather than as a controlled head-to-head experiment.

Table 8: Comparison of the Proposed Framework with Selected Related Studies

| Study               | Method                                       | Dataset / Task                 | Reported Accuracy    |
|---------------------|----------------------------------------------|--------------------------------|----------------------|
| Kumar et al. [28]   | DRISTI: VGG16 + Capsule Network              | APTOS 2019; five-class grading | 75.81% test accuracy |
| Mohanty et al. [23] | VGG16 feature extraction + XGBoost           | APTOS 2019; five-class grading | 79.50%               |
| Islam et al. [29]   | Xception + supervised contrastive learning   | APTOS 2019; five-class grading | 84.36%               |
| Proposed framework  | ConvNeXt-Tiny + Swin-V2-Tiny + Random Forest | APTOS 2019; five-class grading | 84.91%               |

Comparison of the Proposed Framework with Selected Related Studies (APTOS 2019, Five-Class Grading)

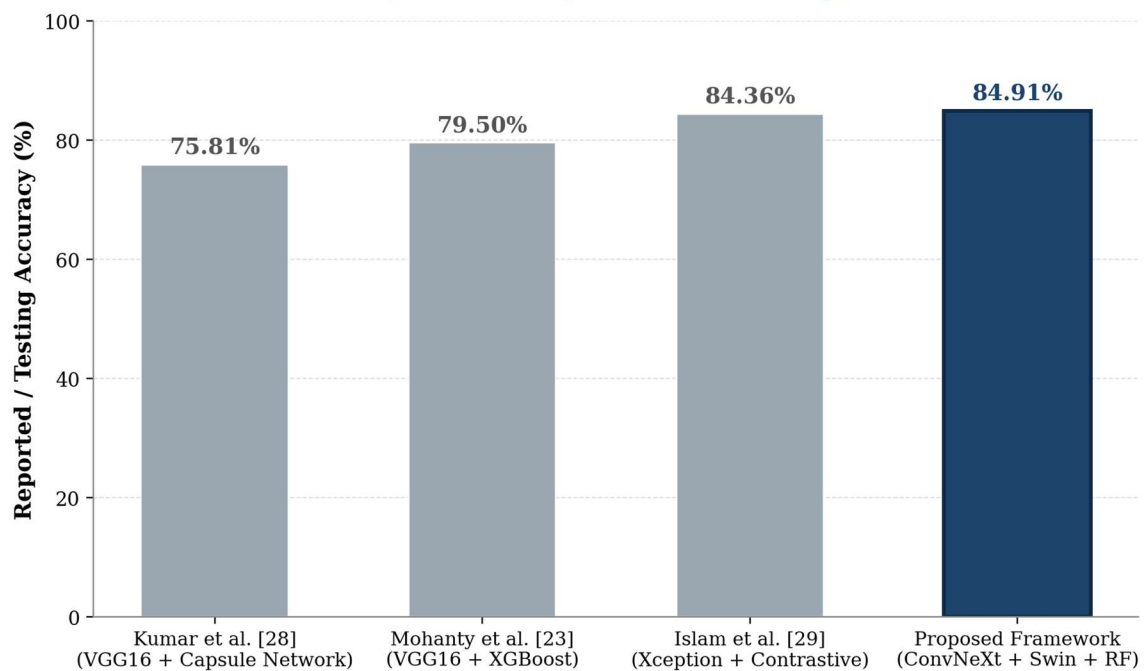


Figure 17: Bar Chart Comparison of the Proposed Framework with Selected Related Studies

## 4.4 Discussion of Findings

### 4.4.1 Contribution of Feature Fusion and Random Forest

The final representation combines two different types of learned information. ConvNeXt-Tiny contributes convolutional features that describe local and hierarchical retinal structure, while Swin-V2-Tiny adds context through shifted-window attention. The gain from fusion was modest rather than dramatic, but concatenating both embeddings still increased the set of predictors available to Random Forest. The ensemble could then model nonlinear relationships among those variables without requiring another end-to-end deep-network training stage.

Training accuracy was noticeably higher than accuracy on held-out images, which shows that the Random Forest fitted the training embeddings more closely than it generalized. To reduce this effect, the selected model used shallow trees, large minimum split and leaf sizes, restricted feature sampling, bootstrap sampling, and balanced-subsample weighting. A 14.04-point difference between training and test accuracy remained, indicating residual overfitting. Validation and test accuracy, however, were separated by only 1.25 percentage points, suggesting that performance was reasonably consistent across the two held-out subsets.

### 4.4.2 Interpretation in Relation to Related Work

At 84.91%, the proposed multiclass accuracy is near the 85.10% result reported by Youldash et al. [18] and is higher than some outcomes reported for IDRiD or other Kaggle-based diabetic retinopathy experiments [22], [27]. Much higher figures appear elsewhere in the reviewed literature, but those studies do not always use the same dataset, number of classes, preprocessing, or partitioning strategy. Consequently, their headline accuracies are not directly comparable with the present protocol.

A strength of the current experiment is the separation of model-development decisions from final testing. The same preprocessing and data split were retained throughout; stochastic augmentation was confined to training; deep checkpoints were chosen from validation results; and Random Forest settings were determined without consulting the test labels. Reporting both aggregate and class-sensitive measures makes the remaining weaknesses, particularly minority-class recall, visible rather than masking them behind a single accuracy value.

## 4.5 Limitations

A remaining limitation is the separation between training performance and the scores obtained on held-out data, most clearly reflected in macro recall and macro F1-score.

Performance remains weakest for the less represented Mild, Severe, and Proliferative DR classes. Improving recall and F1-score for these grades is therefore a priority for future work.

The approach is also computationally demanding because two deep backbones must be fine-tuned and two sets of embeddings must be generated and stored before the classical classifier is trained.

Several Random Forest configurations are then evaluated before the final model is selected, which adds further experimentation time.

## 4.6 Chapter Summary

Chapter 4 evaluated the fused Random Forest on the held-out test set. The final model reached 84.91% accuracy, 70.98% macro F1-score, 84.45% weighted F1-score, 96.11% macro specificity, and 94.80% macro ROC-AUC. The confusion matrix showed that errors were concentrated in adjacent and less represented severity grades. These findings were interpreted in relation to the contribution of feature fusion, residual overfitting, class imbalance, and differences from the reviewed studies.

## Chapter 5 Conclusion and Future Work

This study developed and tested a five-class diabetic retinopathy grading pipeline using the APTOS 2019 Blindness Detection dataset of 3,662 fundus images. The experimental design was built around three practical issues in this task: variation in image acquisition, a strongly uneven class distribution, and visual overlap between adjacent severity grades. A fixed preprocessing sequence was applied to all subsets, while augmentation was restricted to training images so that validation and test data remained unchanged. The model represented each image in two ways. ConvNeXt-Tiny provided convolutional features describing local and hierarchical retinal structure, whereas Swin-V2-Tiny supplied a second embedding that captured wider spatial relationships through shifted-window attention. Each backbone produced 768 features. Their concatenation formed a 1,536-dimensional representation that was classified by a regularized Random Forest. Separating feature extraction from the final classifier also made it possible to save the embeddings and reuse them when comparing Random Forest configurations. Taken together, the experiment indicates that a mixed convolutional-transformer representation can be used successfully with Random Forest for five-class diabetic retinopathy grading. ConvNeXt-Tiny and Swin-V2-Tiny provide different forms of visual encoding, and their concatenation gives the classifier access to both local and hierarchical detail and broader contextual information. The present findings should nonetheless be treated as dataset-specific because validation was performed on APTOS 2019. Future work should therefore test the pipeline on independent retinal datasets and images from different acquisition devices.

Additional work is also needed on minority-class performance, for example through stronger class-balancing strategies that do not compromise evaluation integrity. Finally, adding explain ability methods such as Grad-CAM or other attribution techniques could make model decisions easier to inspect in a clinical context.

## References

- [1] Sebastian, A., Elharrouss, O., Al-Maadeed, S., & Almaadeed, N. (2023). A survey on deep-learning-based diabetic retinopathy classification. *Diagnostics*, 13(3), 345.
- [2] Haq, N. U., Waheed, T., Ishaq, K., Hassan, M. A., Safie, N., Elias, N. F., & Shoaib, M. (2024). Computationally efficient deep learning models for diabetic retinopathy detection: A systematic literature review. *Artificial Intelligence Review*, 57, 309.
- [3] Senapati, A., Tripathy, H. K., Sharma, V., & Gandomi, A. H. (2024). Artificial intelligence for diabetic retinopathy detection: A systematic review. *Informatics in Medicine Unlocked*, 45, 101445.
- [4] Prathibha, S., & Siddappaji. (2024). Advancing diabetic retinopathy diagnosis with fundus imaging: A comprehensive survey of computer-aided detection, grading and classification methods. *Global Transitions*, 6, 93-112.
- [5] Ikram, A., Imran, A., Li, J., Alzubaidi, A., Ahim, S. F., Yasin, A. Y., & Fathi, H. F. (2024). A systematic review on fundus image-based diabetic retinopathy detection and grading: Current status and future directions. *IEEE Access*, 12, 96273-96303.
- [6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.
- [7] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- [8] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

- [9] Bappi, M. I., Juthy, J. A., & Kim, K. (2025). Deep learning-based diabetic retinopathy recognition and grading: Challenges, gaps, and an improved approach - A survey. *ICT Express*, 11(5), 993-1013.
- [10] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976-11986.
- [11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth  $16 \times 16$  words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- [12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022.
- [13] Grzybowski, A., & Jin, K. (2026). Artificial intelligence-based medical devices for diabetic retinopathy screening in the European Union. *Ophthalmology and Therapy*, 15(2), 569-589.
- [14] Jabbar, A., Liaqat, H. B., Akram, A., Sana, M. U., Azpiroz, I. D., de la Torre Diez, I., & Ashraf, I. (2024). A lesion-based diabetic retinopathy detection through hybrid deep learning model. *IEEE Access*, 12, 40019-40036.
- [15] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2022). Swin Transformer V2: Scaling Up Capacity and Resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12009-12019.

- [16] Akhtar, S., Aftab, S., Ali, O., Ahmad, M., Khan, M. A., Abbas, S., & Ghazal, T. M. (2025). A deep learning based model for diabetic retinopathy grading. *Scientific Reports*, 15, 3763.
- [17] Guefrachi, S., Echioui, A., & Hamam, H. (2025). Diabetic retinopathy detection using deep learning multistage training method. *Arabian Journal for Science and Engineering*, 50, 1079-1096.
- [18] Youldash, M., Rahman, A., Alsayed, M., Sebiany, A., Alzayat, J., Aljishi, N., Alshammari, G., & Alqahtani, M. (2024). Early detection and classification of diabetic retinopathy: A deep learning approach. *AI*, 5(4), 2586-2617.
- [19] Shamrat, F. M. J. M., Shakil, R., Sharmin, S., Ovy, N. H., Akter, B., Ahmed, M. S., Ahmed, M. Z., Ahmed, K., Bui, F. M., & Moni, M. A. (2024). An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection. *Healthcare Analytics*, 5, 100303.
- [20] Fayyaz, A. M., Sharif, M. I., Azam, S., Karim, A., & El-Den, J. (2023). Analysis of diabetic retinopathy (DR) based on deep learning. *Information*, 14(1), 30.
- [21] Alwakid, G., Gouda, W., & Humayun, M. (2023). Enhancement of diabetic retinopathy prognostication using deep learning, CLAHE, and ESRGAN. *Diagnostics*, 13(14), 2375.
- [22] Lin, C.-L., & Wu, K.-C. (2023). Development of revised ResNet-50 for diabetic retinopathy detection. *BMC Bioinformatics*, 24, 157.
- [23] Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V. C., Karamitsos, I., & Kanavos, A. (2023). Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23(12), 5726.

- [24] Patil, M. S., Chickerur, S., Abhimalya, C., Naik, A., Kumari, N., & Maurya, S. (2023). Effective deep learning data augmentation techniques for diabetic retinopathy classification. *Procedia Computer Science*, 218, 1156-1165.
- [25] Ali, A. A., & Dawood, F. A. A. (2023). Deep learning of diabetic retinopathy classification in fundus images. *Journal of Engineering*, 29(12), 139-152.
- [26] Das, D., Biswas, S. K., & Bandyopadhyay, S. (2023). Detection of diabetic retinopathy using convolutional neural networks for feature extraction and classification (DRFEC). *Multimedia Tools and Applications*, 82(19), 29943-30001.
- [27] Jiwani, N., Gupta, K., & Afreen, N. (2022). A convolutional neural network approach for diabetic retinopathy classification. In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 357-361). IEEE.
- [28] Kumar, G., Chatterjee, S. K., & Chattopadhyay, C. (2021). DRISTI: a hybrid deep neural network for diabetic retinopathy diagnosis. *Signal, Image and Video Processing*, 15(8), 1679-1686.
- [29] Islam, M. R., Abdulrazak, L. F., Nahiduzzaman, M., Goni, M. O. F., Anower, M. S., Ahsan, M., Haider, J., & Kowalski, M. (2022). Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine*, 146, 105602.