

# اسراء الوتار

## An\_Efficient\_Hybrid\_Framework\_References\_Final\_Reviewe...

 Midocean University

### Document Details

Submission ID

trn:oid::2945:408961606

Submission Date

Aug 21, 2026, 7:10 PM GMT+3

Download Date

Aug 21, 2026, 7:33 PM GMT+3

File Name

An\_Efficient\_Hybrid\_Framework\_References\_Final\_Reviewed.docx

File Size

8.0 MB

61 Pages

9,320 Words

56,944 Characters

## 0% 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.



1 AI-generated 0%

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 one of the main reasons behind avoidable vision loss, but it is not easy to categorize the severity level using retinal images due to issues such as poor image quality, unbalanced classification data, and similarity of the adjacent classes. In this thesis, we analyze a five-class classification approach that uses deep feature extraction along with a classical machine learning classifier. Two pre-trained backbones were fine-tuned on the APTOS 2019 Blindness Detection dataset, ConvNeXt-Tiny to extract localized and hierarchical features of the retina, and Swin-V2-Tiny for context. 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 F1score, 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 (DR) is a progressive microvascular complication of diabetes and an important cause of preventable visual impairment. Prolonged hyperglycemia damages the retinal vasculature and may produce microaneurysms, hemorrhages, exudates, ischemic changes, and neovascularization as the disease progresses [1]. Clinical grading is commonly performed by specialists through visual examination of retinal fundus images; however, manual assessment is time-consuming, depends on reader expertise, and is difficult to scale for population-level screening [2]. Advances in artificial intelligence have therefore encouraged the development of automated systems capable of learning disease-related patterns directly from retinal images [3].

The current research presents a novel five-class diabetic retinopathy grading methodology based on deterministic image processing, data augmentation from training only, the transfer learning techniques like ConvNeXt-tiny and Swin-V2-Tiny, fusion of features, and Random Forest classifications. The technique is developed to merge the hybrid systems of convolutional and transformer representations to classify retinal images into five classes including no DR, mild DR, moderate DR, severe DR, and proliferative DR.

## 1.2 Problem Definition

Diabetic retinopathy grading with automated systems poses challenges because of variation in resolution, illumination, contrast, field of view, and image capturing process. The problem is further exacerbated due to imbalanced data and fine-grained nature of neighboring severity classes especially the severe and proliferative ones which are relatively rare in the dataset. Convolutional Neural Networks perform well in identifying local textures and lesions; however, a single CNN architecture will not effectively represent the spatial relationships present in retinal fundus images. On the other hand, Transformer models can learn interdependencies among various parts of retinal images but might need proper regularization and large amounts of data. Therefore, this thesis explores the possibility that a regularized Random Forest can classify the fusion of deep embeddings generated using ConvNeXt-Tiny and Swin-V2-Tiny. In addition, this thesis utilizes dataset integrity check, stratified split, weight optimization, controlled augmentation, holdout validation, and testing to prevent data leakage.

## 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** reports and discusses the experimental results of the proposed hybrid classifier using the defined evaluation measures and class-level analyses.
- **Chapter 5** In the final chapter of the thesis, we recapitulate the main results, highlight the methodical and practical constraints of this study, and point out new research avenues in the field of explainability as well as clinical validation.

## Chapter 2: Background and Related Work

### 2.1 Introduction

This chapter establishes the medical, computational, and empirical foundations of the proposed diabetic retinopathy classification framework. It first describes the disease and its five severity grades, then reviews the role of artificial intelligence in retinal-image analysis. The discussion covers image preprocessing, data augmentation, transfer learning, machine-learning classifiers, convolutional neural networks, vision transformers, deep embeddings, and feature-level fusion. The chapter also defines the performance measures required for imbalanced multiclass evaluation. Finally, fourteen related studies are critically reviewed and compared to identify the methodological gap addressed by the proposed ConvNeXt-Tiny, SwinV2-Tiny, and Random Forest framework

### 2.2 Background

#### 2.2.1 Diabetic Retinopathy Overview

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus in which prolonged hyperglycemia damages retinal blood vessels. Clinical manifestations include microaneurysms, hemorrhages, hard exudates, cotton-wool spots, venous abnormalities, intraretinal microvascular abnormalities, and neovascularization. The grades of the disease range from no DR to mild, moderate, severe non-proliferative and then ultimately to proliferative DR. Grading is critical to ensure timely treatment of the condition as its severity progresses where visual impairment increases as well. Since the early stages are symptom-free, it is equally important to perform timely screenings and to accurately interpret retinal fundus images. [4].

### 2.2.2 Artificial Intelligence in Medical Image Analysis:

Artificial intelligence (AI) is an increasingly important tool for the analysis of medical images because computer systems can learn complex visual patterns relevant to the identification of pathology, evaluation of severity, and guidance of clinical treatment. Retinal analysis generally follows a fully automated pipeline that consists of image quality control, pre-processing, data augmentation, feature extraction through deep learning, classification, and performance evaluation. Pre-processing includes operations like background removal, contrast enhancement, resizing, and normalization in order to minimize variance in the acquired images while augmentation serves to diversify the training set and reduce bias toward the dominant classes of data. The classical machine-learning paradigm separates feature extraction from classification, while deep-learning techniques are able to learn hierarchical representations from the images. The principle of transfer learning is particularly useful in the processing of medical data due to the paucity of labeled data because neural networks obtained from training on other types of medical images can be employed on images taken of the retina. Hybrid approaches utilize this concept by employing the deep feature representations of different neural networks prior to classification. Convolutional neural networks segregate local textural information from the lesion itself; on the other hand, transformer networks provide information about the context in which the lesions appear. The latter has been combined with the former to provide a deeper understanding of the features of retinal images [4], [5]. The various methods of AI employed in retinal image processing are presented in Figure 1

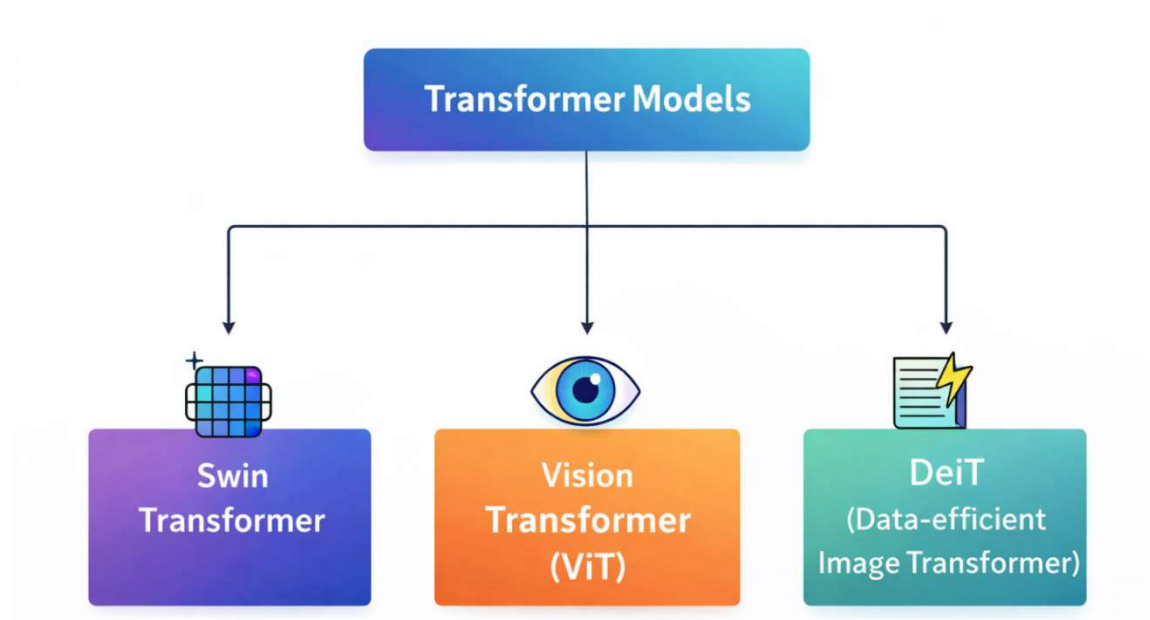


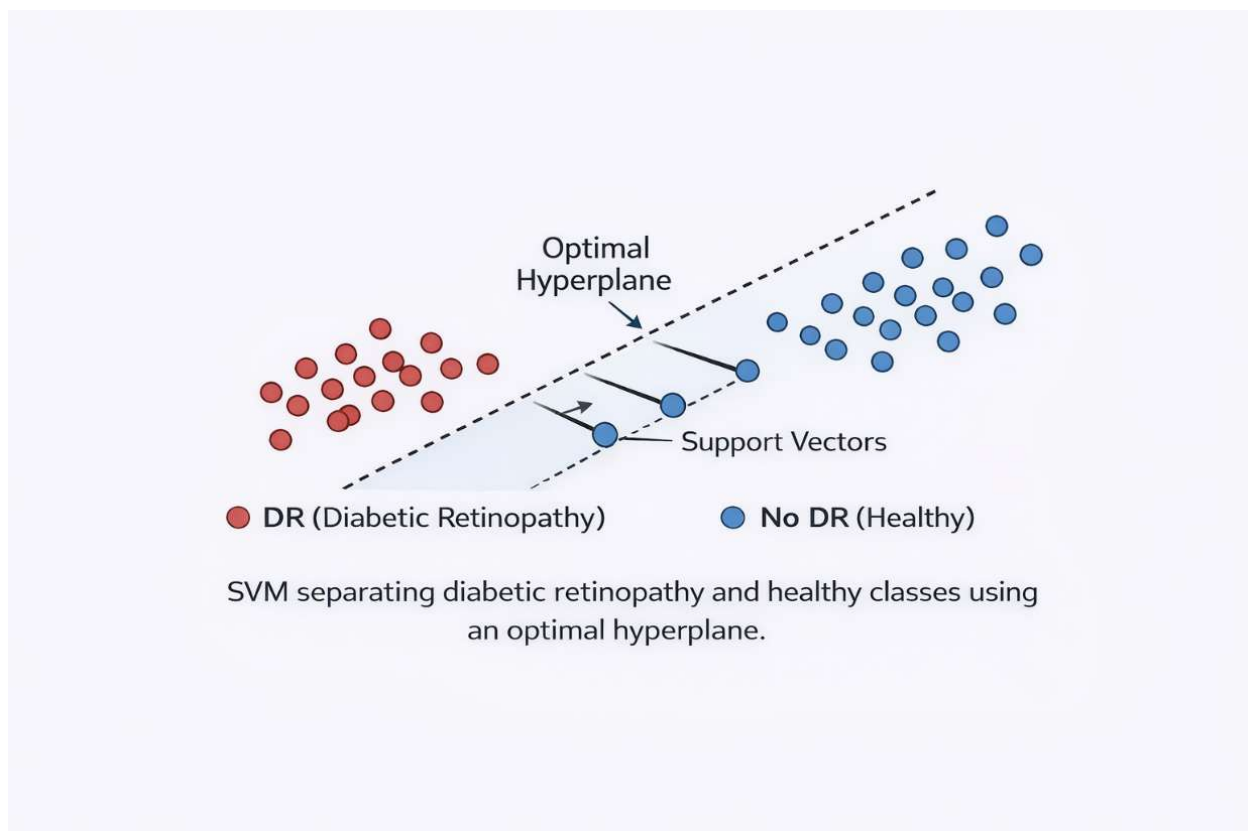
Figure 1: AI Techniques for Retinal Image Analysis [4] [5]

### 2.2.3 Diabetic Retinopathy Detection Using Machine Learning

The analysis of diabetic retinopathy with a machine-learning technique usually consists of two steps: firstly, extracting the features from retinal photographs and secondly, creating a model to perform classification based on these extracted features. Features may be derived manually based on texture and pattern recognition or automatically using models already trained on the specific dataset. The introduction of this separation of processes allows reducing the time and costs associated with the implementation of a complete end-to-end network and permits the testing of different classification algorithms over the same set of images with fixed parameters. The proposed system uses the Random Forest classification model applied to the ConvNeXt-Tiny and Swin-V2-Tiny embeddings that have been generated during the previous stage of the process.

## 1. Support Vector Machine (SVM)

A support vector machine forms a maximum margin hyperplane; it can employ kernels to accomplish nonlinear classification. In retinal image analysis, the support vector machine is used on engineered or deep feature representations, but its efficacy is dependent on scaling, kernel choice, and tuning of hyperparameters [6] as shown at Figure 2.



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

## 2. Random Forest

Random Forest refers to an ensemble classifier that builds up several decision trees on the bootstrapped samples and the randomly chosen feature sets. The final classification is obtained through the aggregation of decisions of the trees. Its non-linear decision-making structure and robustness in the presence of correlated predictors and class-weighting capability render it applicable for classification of unbalanced retina image data and fused deep embedding [7]. Figure 3 shows its mechanism of majority voting.

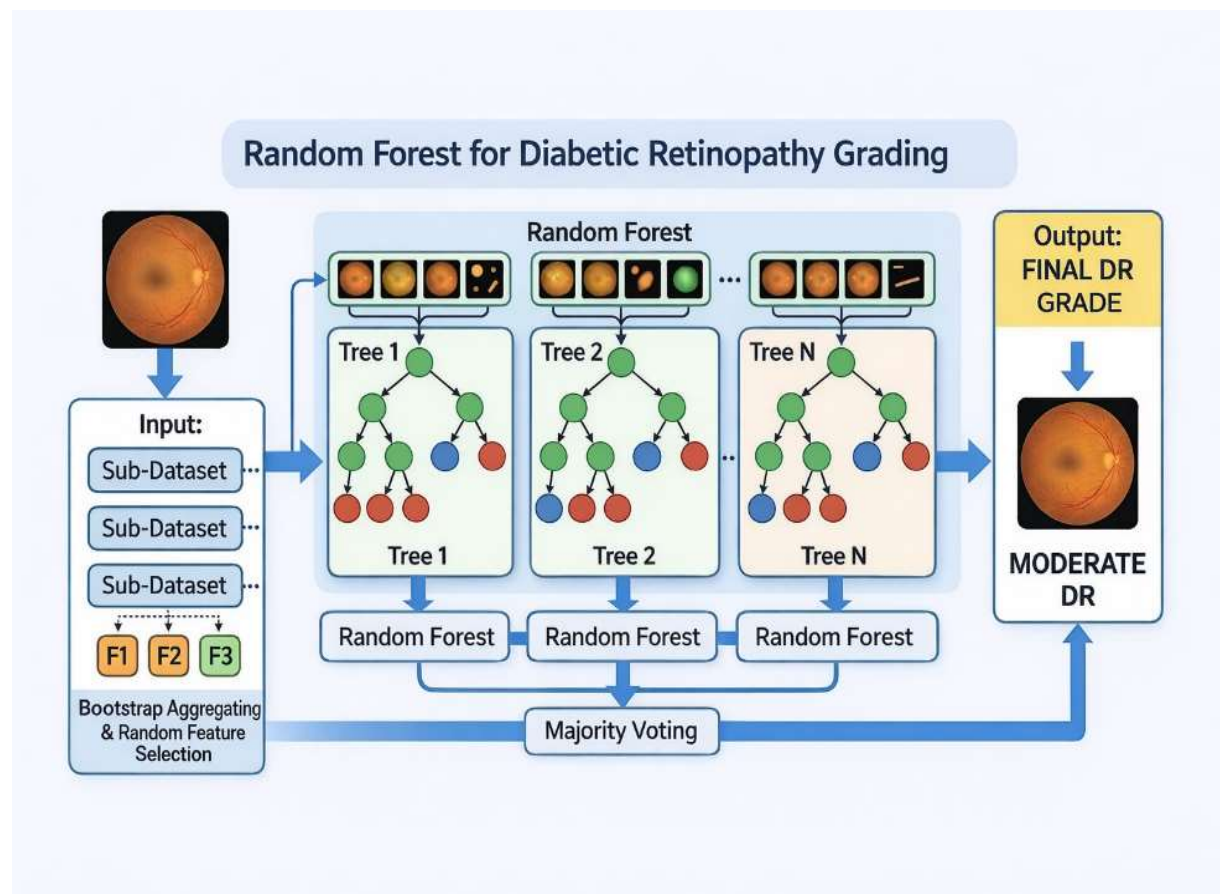


Figure 3: Random Forest Classification for Diabetic Retinopathy Grading [7]

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

K-Nearest Neighbor classifier works on the principle of classifying an object by considering the labels of the closest data samples within the feature space. Its simple, understandable method of neighborhood voting can be helpful to use it as a baseline. Performance of K-Nearest Neighbor depends on distance measure, scaling of features, size of the neighborhood, and dimension of the data. The image shown in Figure 4 depicts the application of K-Nearest Neighbor method for grading of retinal images.

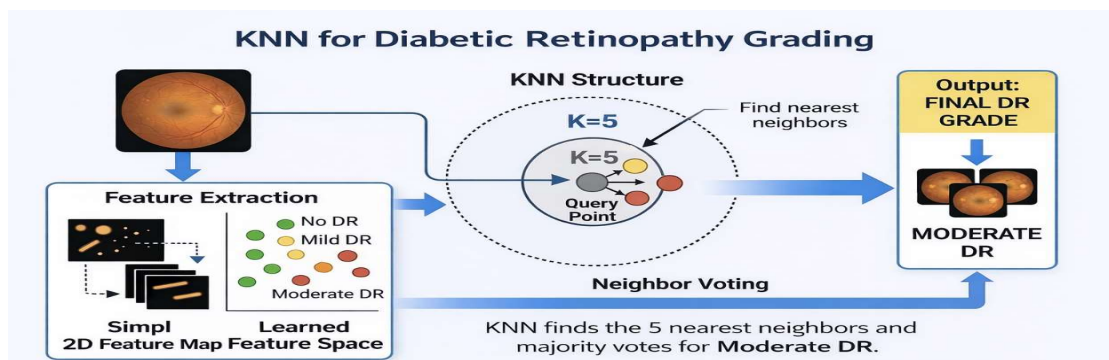


Figure 4: K-Nearest Neighbors for Diabetic Retinopathy Grading [8]

#### 2.2.4 Diabetic Retinopathy Detection Using Deep Learning

Deep learning has transformed diabetic retinopathy analysis by learning hierarchical The adoption of deep learning has revolutionized the analysis of diabetic retinopathy owing to the capacity of this technology to acquire hierarchical visual abstractions from retinal fundus images. Early level layers acquire edge and color change detection as well as fine texturing, while its layers acquire the capacity to identify blood vessels, lesions, and various structures related to illness. However, success in this process is connected to image quality, or class unbalance, all the researches conducted previously, and different necessary criteria that must be taken into

consideration [9]. This study investigates the opportunities provided by the ConvNeXt-Tiny convolutional architectural feature extraction technique as the most modern one, keeping the ability to form effective convolutional representations simultaneously.

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

Through larger kernels and stage-wise feature hierarchy, ConvNeXt-Tiny updates conventional convolution. It allows for characterizing local texture of lesions, vascular organization, and more abstract diseased patterns in retinal images. In the suggested methodology, fine-tuning has culminated in obtaining a 768-dimensional embedding after the global average pooling stage. After this, the classification head has been dropped and embedding remains for further synthesis [10]. An original conceptual representation of this extraction process can be seen in Figure 5.

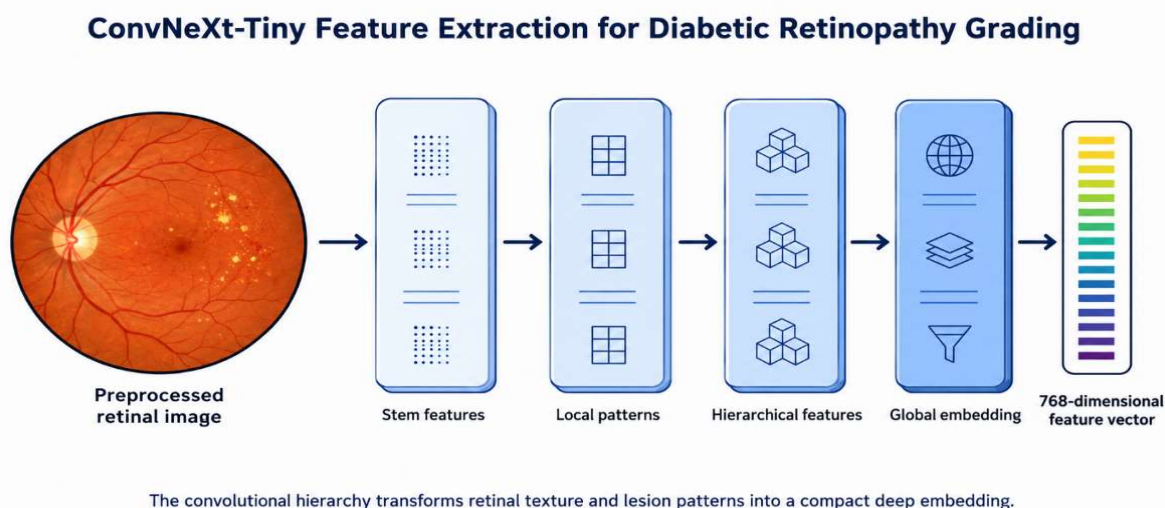


Figure 5: ConvNeXt-Tiny Feature Extraction for Diabetic Retinopathy Grading [10]

## 2.2.5 Diabetic Retinopathy Detection Using Transformer Models

Representation of the image in the vision transformer approach can be viewed as temporal sequence of patches. The vision transformer predicts relationships among them through the self-attention mechanism. Therefore, unlike convolutional models, which aim mainly at identifying local patterns, the construction of vision transformers ensures capturing relations between the areas of the retina that are far apart from each other. However, successful implementation of the method requires regularizing mechanisms, sufficient amount of data as well as considerable computational power. The main representative models based on transformers are shown in figure 6.

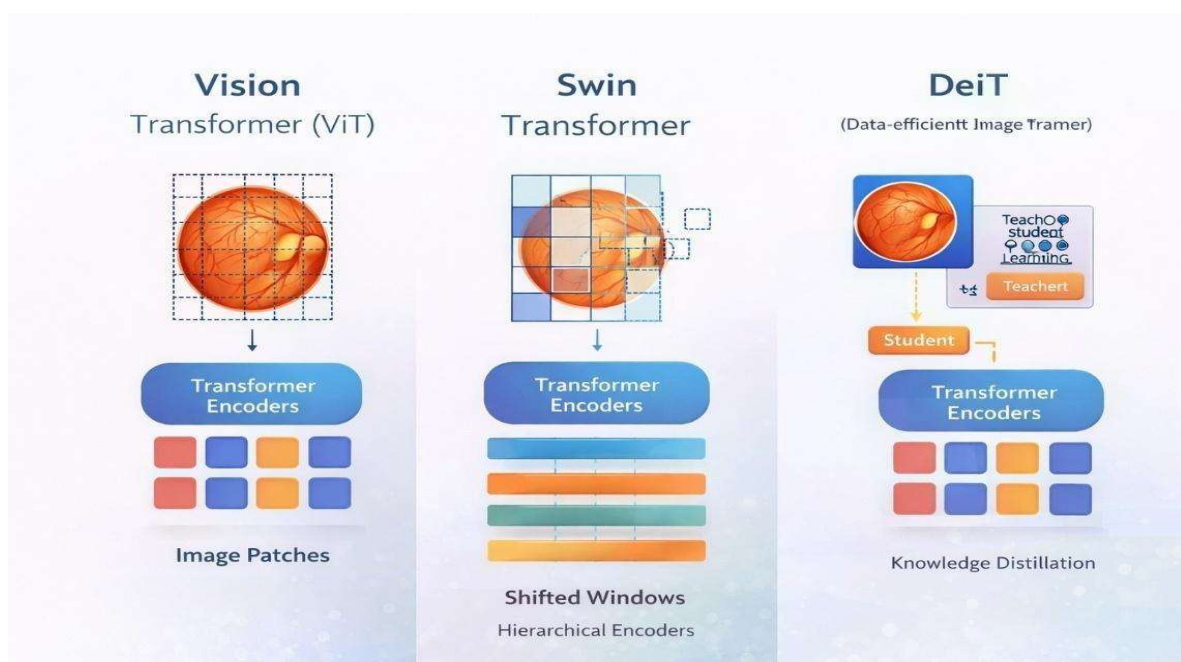


Figure 6: Transformer-Based Retinal Image Encoding [11] [12]

## 2- 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 [11]. Figure 7 illustrates patch-based transformer encoding applied to retinal image analysis.

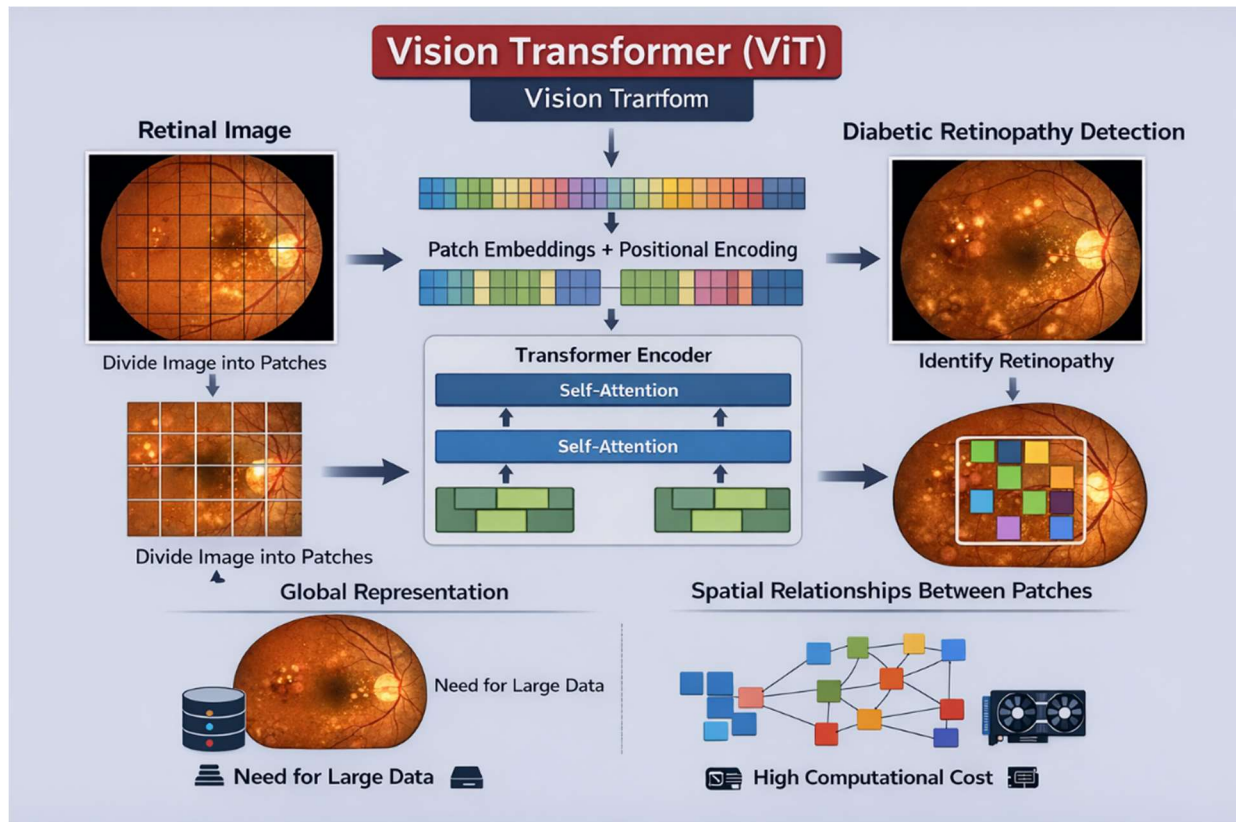


Figure 7: Vision Transformer Encoding for Diabetic Retinopathy Analysis [11]

### 3- Swin Transformer

The Swin Transformer model implements self-attention in confined areas known as windows and updates the partition of these windows in consecutive blocks. This hierarchical technique results in savings in computation as it allows the flow of information to occur across neighboring fields on various spatial scales. Thus, it is an appropriate technique for the analysis of retinal data since both small lesions and wide context patterns can be analyzed and represented [12]. The process of shifted-window feature encoding is demonstrated in Figure 8.

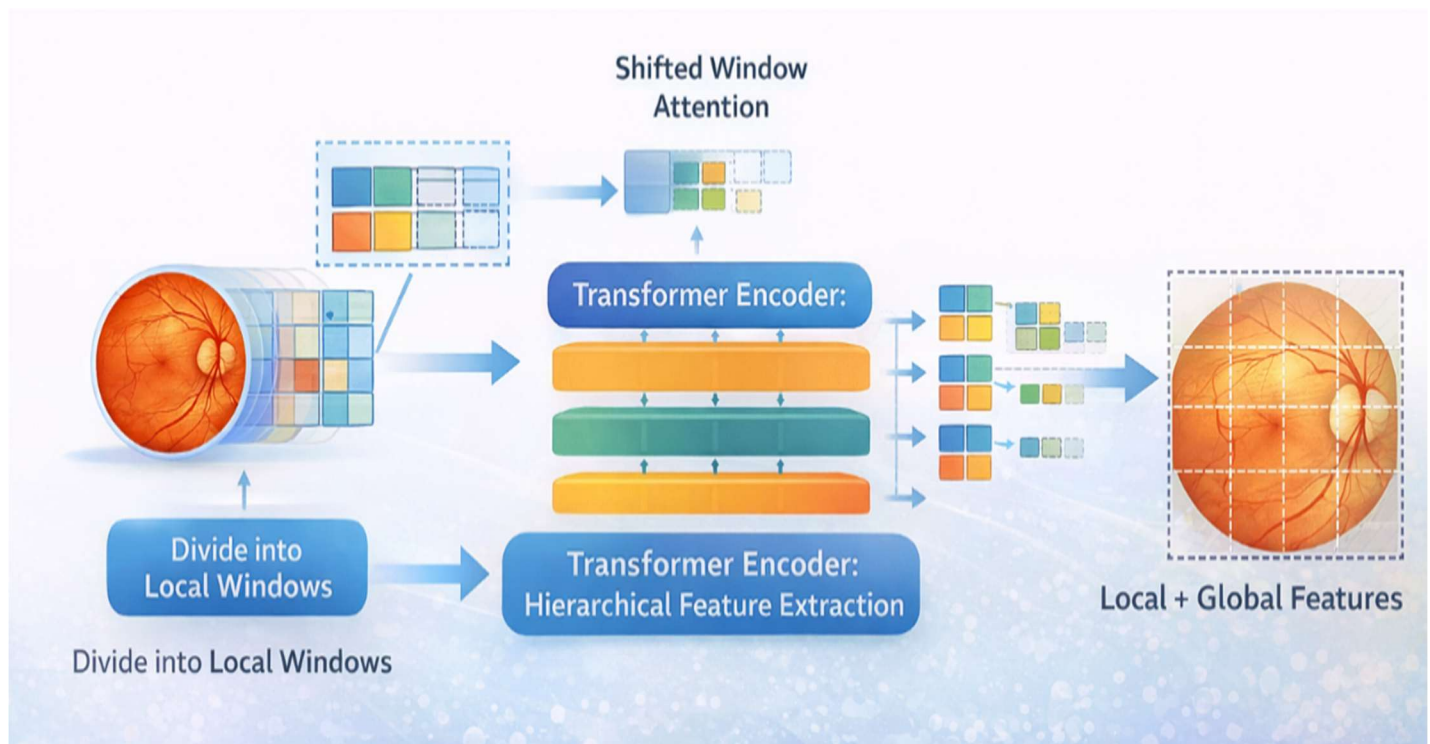


Figure 8: Swin Transformer Encoding for Diabetic Retinopathy Grading [12]

## 2.2.6 Deep Feature Fusion and Hybrid Classification

Feature-level fusion combines representations extracted by different architectures before classification. In retinal-image analysis, ConvNeXt-Tiny can emphasize local texture, while Swin-V2-Tiny performs much broader contextual relationship encoding thanks to shifted-window attention. 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], [12]. 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].

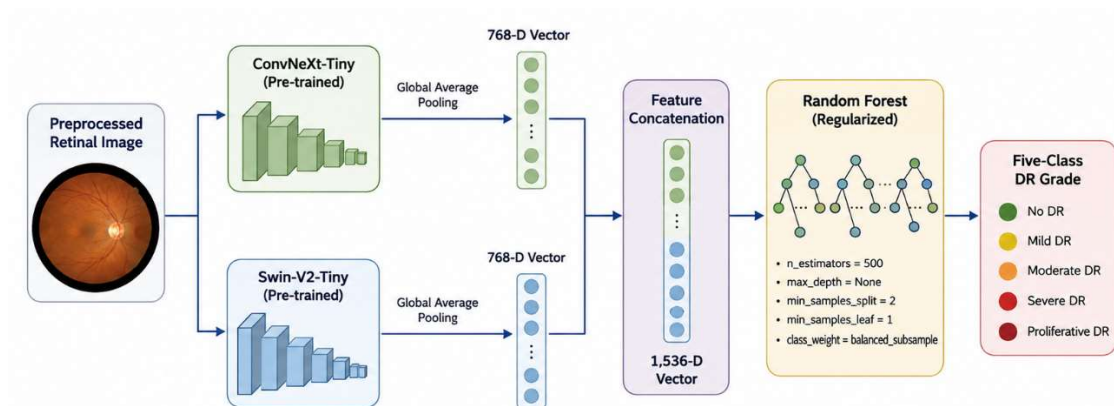


Figure 9: ConvNeXt-Swin Feature Fusion and Random Forest Classification [7]  
[10] [12]

### 2.2.7 Performance Metrics

Evaluation of diabetic retinopathy classifiers requires more than overall accuracy because the clinical classes are commonly imbalanced and errors have different implications across severity grades. A model can achieve a high accuracy while performing poorly on rare severe cases. Accordingly, this thesis considers accuracy together with per-class precision, recall, F1-score, specificity, confusion matrices, and one-vs-rest ROC-AUC values. Macro averages are used to give equal importance to each class, whereas weighted averages reflect the observed class distribution.

- **Accuracy**

Accuracy refers to the ratio of correct classifications to total samples. Where:

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

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

- **Precision**

Precision measures the proportion of correctly predicted positive cases out of all predicted positive cases.

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

- **Recall (Sensitivity)**

Recall refers to the proportion of actual positive cases that are successfully predicted by the model.

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

- **F1-Score**

The F1-score is the definite mean of precision and recall.

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

- **Specificity**

Specificity is defined as the measurement metric to evaluate the model's ability to point out correct negative instances.

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

- **Confusion Matrix**

Confusion matrix represents the results of classification in terms of TP, TN, FP and FN. It helps to indicate the types of error committed by the model.

- **ROC Curve and AUC**

Receiver Operating Characteristic curve depicts the trade-off between sensitivity and specificity at various threshold values. Area Under the Curve (AUC) signifies the model's capacity to distinguish classes.

- **AUC = 1** represents excellent classification.
- **AUC = 0.5** represents random classification.

The aforementioned measurements are very important in case of understanding five-class diabetic retinopathy grading, since it helps to identify if performance was balanced in terms of non-diabetic retinopathy, mild, moderate, severe and proliferative diabetic retinopathy or it was imposed by the overrepresented class.

## 2.3 Literature Review

Grzybowski and Jin [13] performed a targeted scoping review of 13 artificial intelligence systems with CE certification that had to do with autonomous or semi-autonomous diabetic retinopathy screening in European Union countries. The analysis showed that deep learning based analysis of the fundus images is prevalent among different systems, however the validation processes employed, practices in reporting, imaging devices utilized, and characteristics of the population studied were widely different across various technologies. Hence, the authors could not pool data together to assess their diagnostic performance or to rank the systems. The information on sight-affecting diabetic retinopathy was not very systematic, with the mean metrics hovering around 80% across all controlled evaluations related to the analyzed devices. The study presents value since it illustrates the challenges related to the regulatory and practical implementation issues that arise from applying high-end technology; however, the authors were looking at existing commercial systems instead of developing multiclass assessments- an experimental model.

Akhtar et al. [14] suggested a model called RSG-Net intended for the purpose of conducting the four-stage diabetic retinopathy classification; they based their findings on the Messidor-1 dataset. The model includes a set of techniques, such as box cropping, denoising, histogram equalization, resizing, data augmentation, and other measures aimed at optimizing convolutional model architecture. The RSG-Net

model proved to be capable of achieving accuracy metrics of 99.36% pertaining to the four-stage model and 99.37% regarding the binary classification of diabetic retinopathy. Despite being such impressive results, however, the near-ideal training output indicates that there was a high risk of overfitting. Also, the authors have only utilized one dataset, and thus there is no evidence of successful implementation of the model in different cases.

Guefrachi et al. [15] examined a multi-stage training technique built on several well-known convolutional architectures such as VGG16, VGG19, InceptionResNetV2, EfficientNetV2L, MobileNetV2, and DenseNet121. Data augmentation techniques were employed to enhance the robustness of the model and minimize overfitting effects. As a result, the proposed model attained certain accuracy indicators of approximately 96.61% in diabetic retinopathy identification and detection. Although the findings confirm the results of transfer learning, the small size of the dataset limits the generalizability of the results obtained.

Youldash et al. [16] compared six pre-trained CNN families in their experiments on retinal images from a clinical center in Saudi Arabia and APTOS images. For APTOS images, RegNetX080 achieved binary classification accuracy of 98.6% while EfficientNetB3 earned 85.1% for multi-class classification. External clinical results featured a maximum of 98.2% and 84.4% for binary and multi-class classifications, respectively. The experiment proves that binary identification poses significantly fewer challenges than five-class classification and emphasizes the impact of class imbalance and domain shift.

Shamrat et al. [17] introduced a custom convolutional neural network called DRNet13 tailored to five-class diabetic retinopathy classification and compared it to fifteen pre-trained CNN architectures. The APTOS images underwent certain improvements such as gamma correction and median filtering. As a result, the data

augmentation efforts increased the dataset from 3,662 images to 7,500 images. The accuracy of DRNet13 equated to 97% while the algorithm appeared computationally efficient against the backbones utilized. However, the outcomes were obtained from a single publicly available dataset, and lack of independent clinical verification impacts the validity of conclusions regarding the robustness of deployment.

Jabbar et al. [18] suggested a hybrid lesion-based approach where the representations obtained via GoogleNet and ResNet networks were optimized and utilized for classification using Random Forest, SVM, Linear Regression, and Decision Tree algorithms. The success of this hybrid framework is evidenced by 94% accuracy across the benchmark fundus data set. Although this research supports the relevance of deep feature extraction for machine learning classification, it is evident that this study fails to provide satisfactory evidence in favor of cross-dataset verification.

According to Fayyaz et al. [19], feature extraction can be done using AlexNet and ResNet101 deep learning models. Upon feature extraction, the two models were applied together for representation as well as informative variable selection via a special algorithm from the ant-colony optimization system. The results were then assessed using different types of SVM kernels with a total of 750 features selected from the previous steps. The cubic SVM model was able to achieve 93.0% accuracy with corresponding precision, recall, and F1 score values of 93.0%, 93.5%, and 93.2%. It shows that despite the fact that the research describes the potential for deep-feature fusion technique together with conventional classification, it also illustrates limitations related to feature selection methods and evaluation conditions that may make reproducing the research results difficult.

In the study by Alwakid et al. [20], DenseNet121 was utilized on the APTOS 2019 data set for testing the effect of various image-enhancing techniques such as histogram equalization, CLAHE, and ESRGAN on the results of classification of

images. With the use of the CLAHE and ESRGAN technique the accuracy of the model in regards to the 5-class test set was equal to 98.36% with the model being also able to reach a top-2 and top-3 accuracy of 100%. The results of the research prove the idea that contrast enhancement can make the retinal image classification network work better in comparison with a simple retinal image classification network; however, the method relies a lot on complicated image-enhancing processes and has no references to the validation of the developed method with the help of other datasets not processed in the current research.

In the study by Lin and Wu [21], a new architecture of ResNet50 along with the standardized preprocessing of images and development of adaptive learning rate control parameter was proposed. The new version of the model demonstrated its effectiveness based on the reached results: the training accuracy of the model was equal to 83.95%, while the test accuracy was equal to 74.32%. The new architecture has certain advantages in terms of stability of the learning process if compared to other CNN architectures tested in the current study. However, the difference between training and testing conditions shows limitations of the approach related to the fact that managing overfitting process is complicated and the reliability of multiclass generalization can contain certain weaknesses.

The research conducted by Mohanty et al. [22] was devoted to the comparison of a hybrid VGG16-XGBoost model and DenseNet121 that were applied to the APTOS 2019 dataset after data preprocessing and class balancing. As a result, the VGG16-XGBoost achieved an accuracy rate of 79.50%, while DenseNet121 achieved the result of 97.30%, thus proving that a hybrid design cannot guarantee the high levels of prediction accuracy. The research demonstrates the significance of the feature extraction variants and features combining methods applied in the current study; however, the researchers did not validate the proposed solutions.

Finally, Patil et al. [23] tested the ResNet50 architecture on the APTOS dataset and EyePACS dataset after utilizing the preprocessing and augmentation of images. The algorithm has achieved the number of 97.87% with the use of quadratic weighted kappa of 0.985. Cross-dataset evaluation is among the strengths of the undertaken research, yet, reliance on extensive augmentation processes still exist.

In the study presented by Ali and Dawood [24], the authors have trained, tested, and implemented a pre-trained ResNet50 model to categorize the images from APTOS 2019 fundus images for diabetic retinopathy detection. The system designed by the authors has been described as including pre-processing, transfer learning, checkpoints selection, and evaluation in multiple classes. The final results revealed that the accuracy was equal to 98.3%, precision – to 98.4%, recall – to 98.4%, and F1 score – to 98.5%. Given that the authors show remarkable results in learning based on one dataset, it should be emphasized that efficiency of the system cannot be verified without external validation of data.

Das et al. [25] developed a benchmark system DRFEC and tested 26 pre-trained CNN architectures on EyePACS data in order to obtain data extraction and classification. The best performance in validation was exhibited by EfficientNetB4 (throughput of 79.11% in validation and over 99% in training process), while other large networks have demonstrated some cases of overfitting. The comparison done by the authors is of great importance due to showing inefficiency of large training data regarding its effectiveness in obtaining reliable predictions.

Jiwani et al. [26] focused on the use of convolutional neural network in terms of diabetic retinopathy classification through IDRiD dataset. The current study revealed approximately the same efficacy of 77.6%, determining it as a medium result for a relatively small retina dataset.

## 2.4 Comparative Analysis of Related Work

The fourteen studies examined revealed a series of common elements among them. The results show that the importance of pre-processing and augmentation processes cannot be ignored due to a significant influence of acquisition variability and class imbalance on testing results. Second, although using pretrained CNNs helps obtain good results within the same dataset, an evaluation across datasets will lead to worse performance in multiclass prediction and highlight some challenges in terms of generalization. Third, hybrid methods of building models that use deep learning representations and machine learning classifiers seem to be effective, but only if relevant features are extracted (i.e., the choice of models is appropriate and data leakage is avoided). The results demonstrate that binary classification is easier than multi-class severity scoring because a high accuracy result maybe based largely on a poor classification of minority classes. There is only limited literature that discusses features-level fusion involving contemporary convolutional architecture and hierarchical vision transformer with the use of ensemble classifiers. This justifies the combination of ConvNeXt-Tiny and Swin-V2-Tiny embeddings and their classification by means of Random Forest and stratified training, validation, and testing.

*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. [14]      | 2025 | RSG-Net                                   | Messidor-1                      | 99.36% four-stage; 99.37% binary | Single dataset; overfitting risk           |
| Guefrachi et al. [15]   | 2025 | Multistage transfer learning              | Kaggle fundus images            | 96.61% accuracy                  | Limited data; no broad external validation |
| Youldash et al. [16]    | 2024 | EfficientNet and RegNet transfer learning | APTOS and Saudi clinical images | 98.6% binary; 85.1% multiclass   | Lower multiclass performance               |
| Shamrat et al. [17]     | 2024 | DRNet13 vs. 15 pretrained CNNs            | APTOS 2019                      | 97% accuracy                     | Single public dataset                      |
| Jabbar et al. [18]      | 2024 | GoogleNet-ResNet + APSO + ML              | Fundus dataset                  | 94% accuracy                     | Limited external validation                |
| Fayyaz et al. [19]      | 2023 | AlexNet-ResNet101 fusion + SVM            | EyePACS                         | 93.0% accuracy; 93.2% F1-score   | Complex feature-selection pipeline         |
| Alwakid et al. [20]     | 2023 | DenseNet121 + CLAHE/ESRGAN                | APTOS 2019                      | 98.36% five-class accuracy       | Heavy enhancement; single dataset          |
| Lin and Wu [21]         | 2023 | Revised ResNet50                          | Kaggle DR images                | 83.95% train; 74.32% test        | Remaining train-test gap                   |
| Mohanty et al. [22]     | 2023 | VGG16-XGBoost vs. DenseNet121             | APTOS 2019                      | 79.50% vs. 97.30%                | No external validation                     |

|                     |      |                             |                  |                                |                                           |
|---------------------|------|-----------------------------|------------------|--------------------------------|-------------------------------------------|
| Patil et al. [23]   | 2023 | ResNet50 + augmentation/TTA | APTOS to EyePACS | 97.87%; QWK 0.985              | Extensive augmentation                    |
| Ali and Dawood [24] | 2023 | Fine-tuned ResNet50         | APTOS 2019       | 98.3% accuracy; 98.5% F1-score | No external validation                    |
| Das et al. [25]     | 2023 | DRFEC: 26 pretrained CNNs   | EyePACS          | Best validation : 79.11%       | Overfitting and high train-validation gap |
| Jiwani et al. [26]  | 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 suggested architecture utilizes convolutional and Transformer models' visual representations. ConvNeXt-Tiny provides the strong local inductive bias in regard to retinal textures, vessels, exudates, hemorrhages, lesions, etc. Swin-V2-Tiny enhances the features above using hierarchical shifted window self-attention mechanism and capturing more global spatial relationships and context dependencies from fundus images. The pre-last layer of size 768 is saved in both backbones resulting in two complementary embeddings for every fundus 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 heldout testing subset. The complete architecture of the proposed framework is summarized in Figure 10.

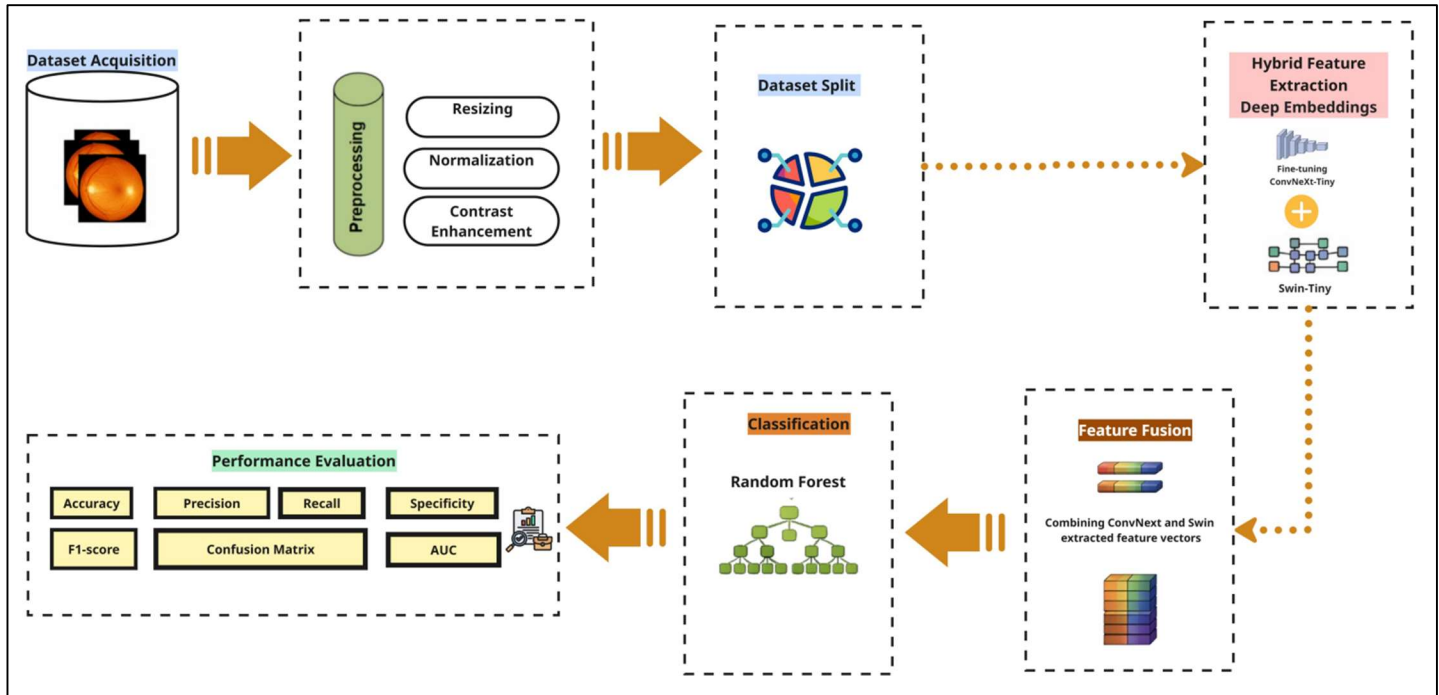
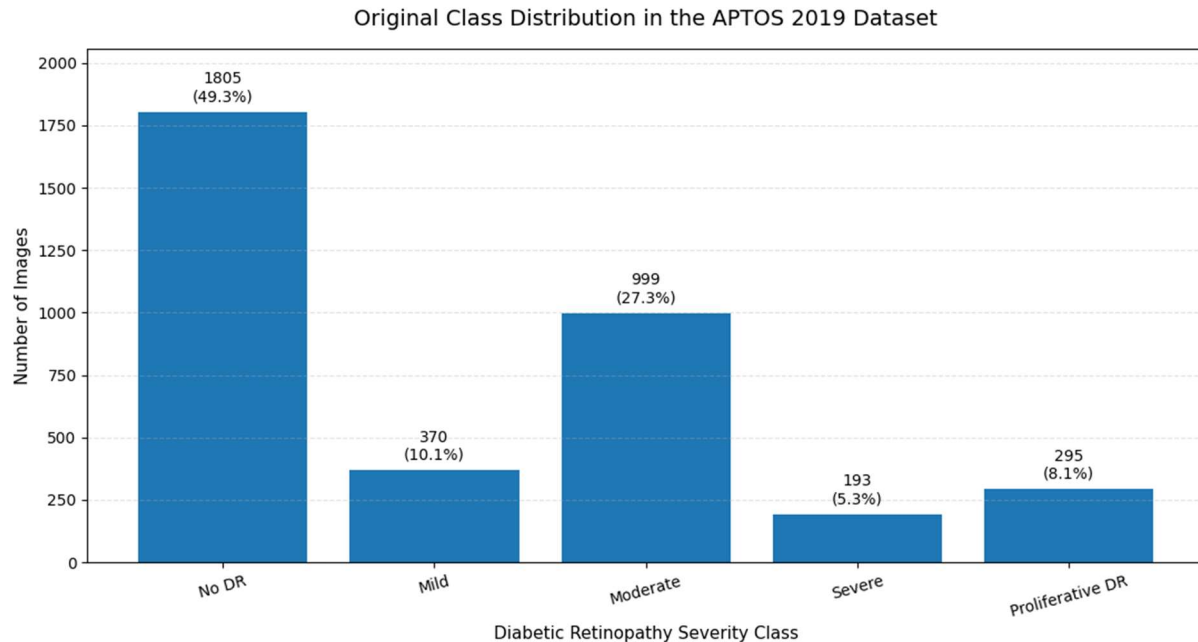


Figure 10: Proposed Hybrid Framework for Diabetic Retinopathy Classification

### 3.3 Dataset Description

For this work, APTOS 2019 Blindness Detection data set is used, which consists of 3,662 color fundus images divided into five classes namely No DR, Mild, Moderate, Severe, and Proliferative DR. Prior to any model training, there was reconciliation of the image files and the label information. This involved checks on missing images, duplicate identifiers, out-of-range labels, and filename label mismatch. 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*

An illustration of one original image picked from each category prior to preprocessing is shown in Figure 12. These images have been selected to show the diversity in the raw dataset in terms of image sizes, lighting, dark edges, retinal field of view, etc.



*Figure 12: Representative Retinal Images Before Preprocessing*



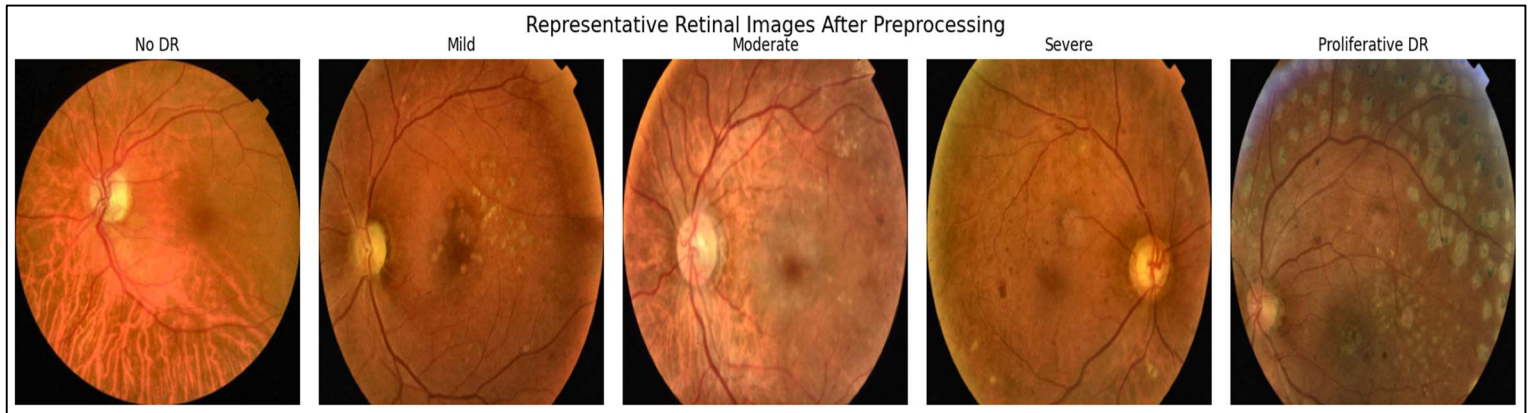
### 3.4 Data Preprocessing

An identical deterministic data preprocessing pipeline was applied to all training, validation, and testing subsets. Stochastic transformations occurred only after preprocessing and only on the training subset. The purpose of deterministic data preprocessing was to decrease variability independent of disease severity and retain retinal structures necessary for grading. Foreground mask was created based on the gray scale version of the image, which served to crop dark periphery areas. The image was then transformed into the LAB color space and CLAHE was applied to the lightness channel to increase local contrast without introducing excessive noise. Contrast-enhanced images were rescaled to 300 x 300 pixels using area interpolation, transformed to RGB float32 PyTorch tensor and normalized using ImageNet channel mean and standard deviation expected by the pre-trained backbones. Table 2 provides the exact values used in the implementation.

*Table 2: Deterministic Retinal-Image Preprocessing Parameters*

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

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

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

### 3.6.2 Swin-V2-Tiny

Swin-V2-Tiny was fine-tuned as the complementary transformer feature extractor [27]. 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

| <i>Parameter</i>          | <i>Value</i>            |
|---------------------------|-------------------------|
| <i>Pretrained weights</i> | <i>ImageNet-1K V1</i>   |
| <i>Input size</i>         | <i>300 x 300 pixels</i> |

|                               |                                           |
|-------------------------------|-------------------------------------------|
| <i>Batch size</i>             | 32                                        |
| <i>Output classes</i>         | 5                                         |
| <i>Embedding dimension</i>    | 768                                       |
| <i>Classifier head</i>        | <i>Dropout (0.20) + Linear (768 to 5)</i> |
| <i>Loss function</i>          | <i>Class-weighted cross-entropy</i>       |
| <i>Label smoothing</i>        | 0.02                                      |
| <i>Backbone learning rate</i> | $1 \times 10^{-5}$                        |
| <i>Head learning rate</i>     | $2 \times 10^{-4}$                        |
| <i>Optimizer</i>              | <i>AdamW</i>                              |
| <i>Best checkpoint epoch</i>  | 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*

| <i>Subset</i>            | <i>Number of Images</i> | <i>Percentage</i> |
|--------------------------|-------------------------|-------------------|
| <b><i>Training</i></b>   | 2,563                   | 70%               |
| <b><i>Validation</i></b> | 549                     | 15%               |
| <b><i>Testing</i></b>    | 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*

| <i>Metric</i>                   | <i>Value</i> |
|---------------------------------|--------------|
| <b><i>Testing accuracy</i></b>  | 0.8491       |
| <b><i>Macro precision</i></b>   | 0.7393       |
| <b><i>Macro recall</i></b>      | 0.6899       |
| <b><i>Macro F1-score</i></b>    | 0.7098       |
| <b><i>Weighted F1-score</i></b> | 0.8445       |
| <b><i>Macro specificity</i></b> | 0.9611       |

|                                |               |
|--------------------------------|---------------|
| <b><i>Macro ROC-AUC</i></b>    | <i>0.9480</i> |
| <b><i>Weighted ROC-AUC</i></b> | <i>0.9684</i> |

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 pattern centers on levels of severity which are adjacent to each other and not by making big jumps from the lowest to the highest level.

Most of the mistakes happened between grades which are close to one another. Mild images are often classified as Moderate images, while Proliferative DR images are often classified as Moderate images too. There are two reasons behind this trend. First, there are fewer images in the minority classes, and second, manifestations of diabetic retinopathy develop slowly and not abruptly between adjacent grades. Moreover, every fundus image is accompanied by one global grade only, and hence the learning objective does not indicate the retinal area containing the lesion evidence.

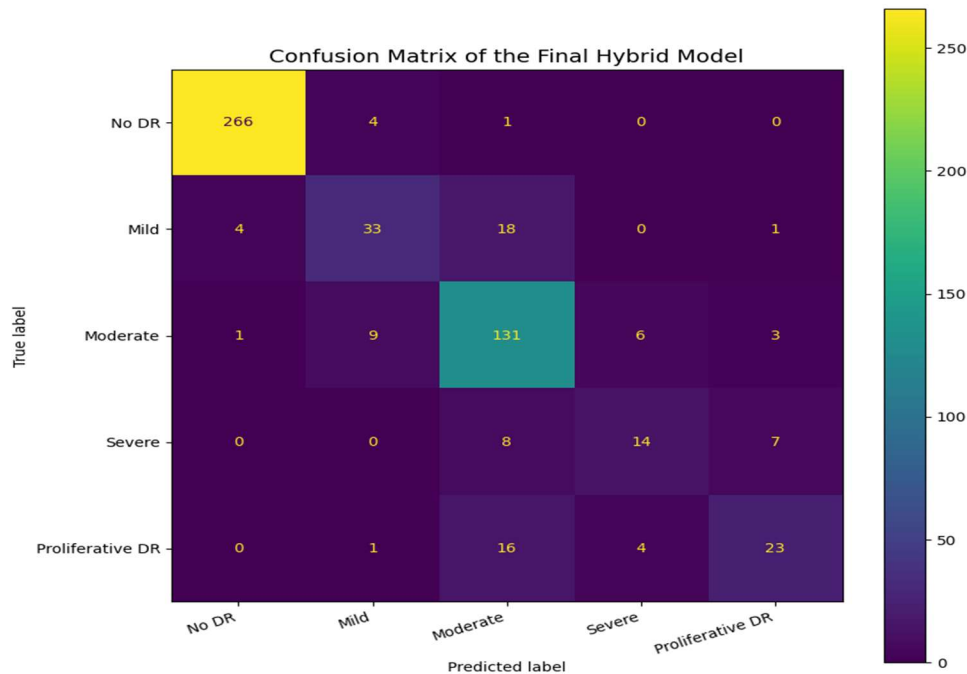


Figure 14: Confusion Matrix of the Final Hybrid Model

Figure 15 presents ROC curves for one-vs-rest obtained using the Random Forest probability predictions. Their AUCs were 0.9963, 0.9271, 0.9506, 0.9340, and 0.9321 for No DR, Mild, Moderate, Severe, and Proliferative DR, respectively. All of the curves stay clearly above the random line. This means that the probability values usually put positive examples above negative ones even if the recall at the decision threshold is not high for some of the classes.

The macro ROC-AUC for the five classes is 0.9480, but after applying frequency weights, it is increased to 0.9684. This is due to the fact that the two most frequent classes, No DR and Moderate, have good ranking properties. Therefore, ROC-AUC is considered along with recall, F1-score, specificity, and the confusion matrix.

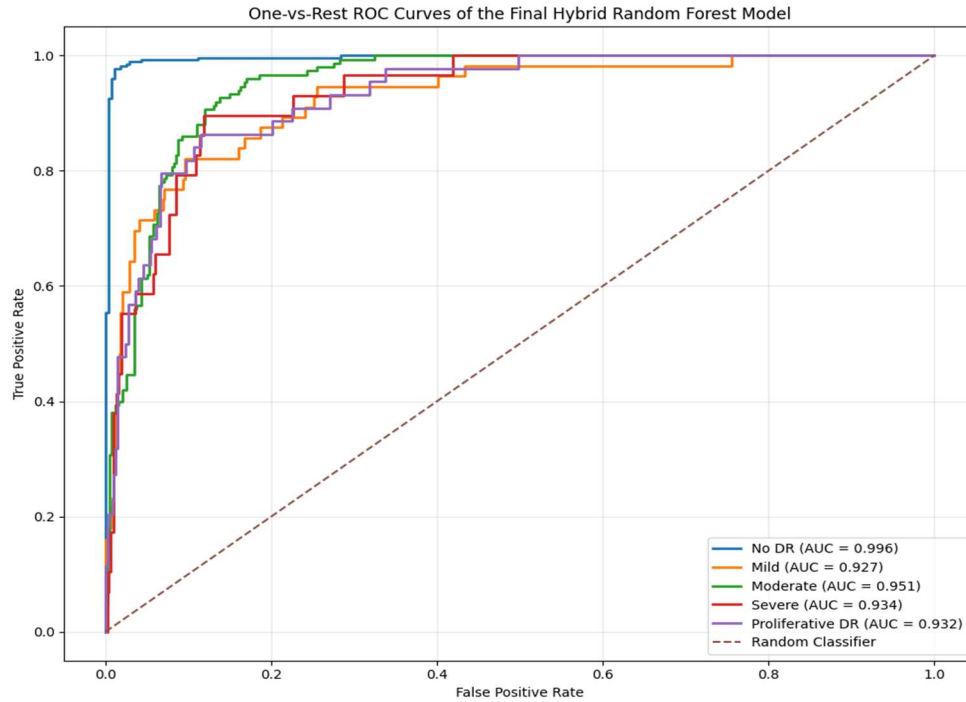


Figure 15: One-vs-Rest ROC Curves of the Final Hybrid Model

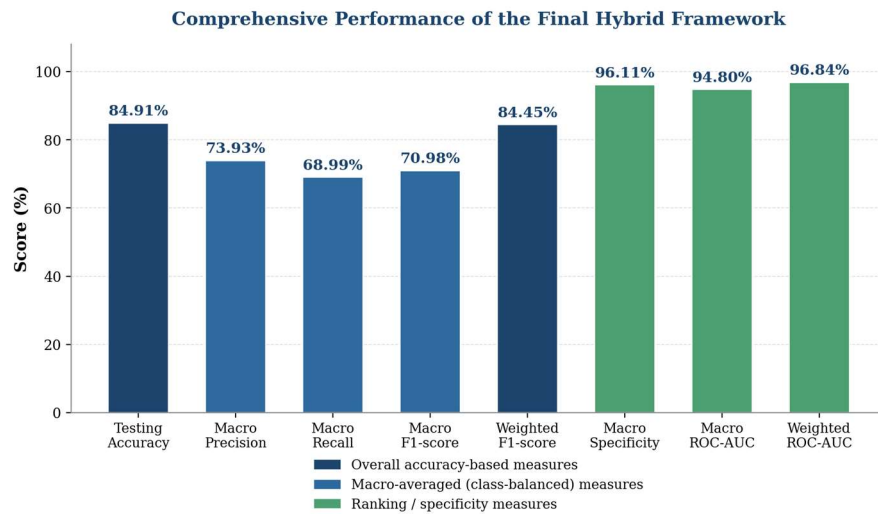


Figure 16: Bar Chart of the Comprehensive Performance Metrics of the Final Hybrid Framework

Table 8 provides a comparative analysis of the developed method against some other studies using APTOS 2019 for five-class diabetic retinopathy assessment. It may be

said that restricting the comparison to one and the same dataset and five classes makes the findings more relevant, but still it is not possible to make an equal comparison since there are differences in other aspects like data preprocessing, partitioning, augmentation, feature extraction, etc.

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. [22] | 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%               |

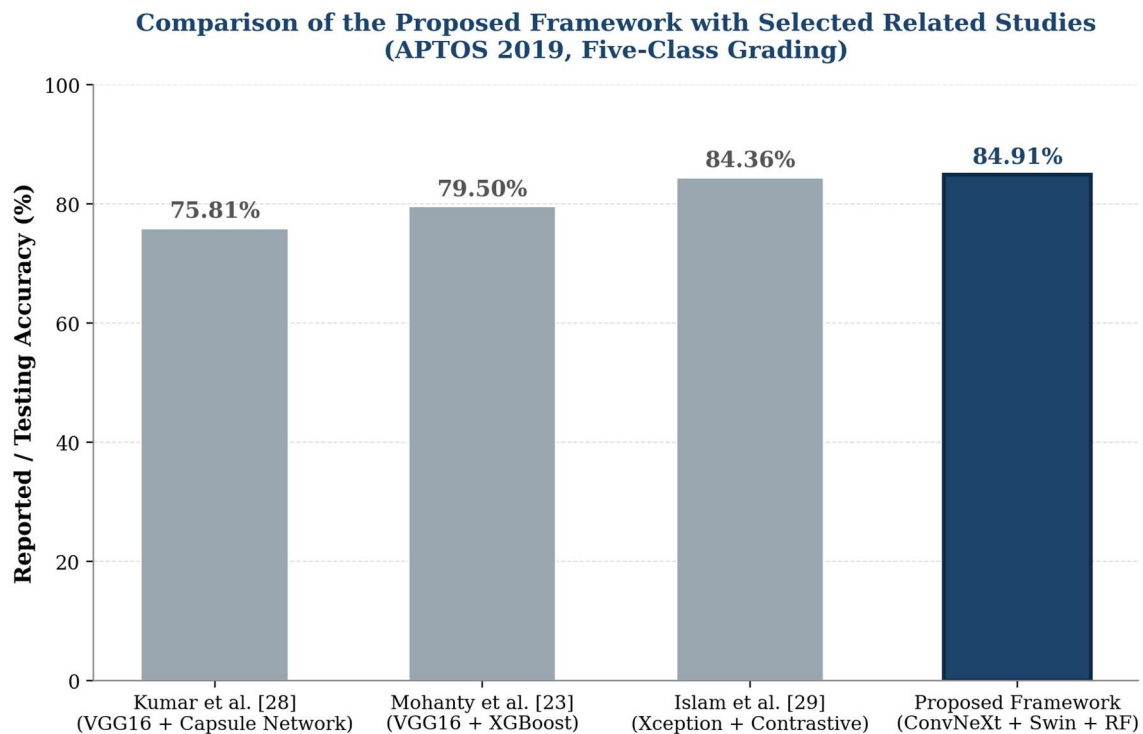


Figure 17: Bar Chart Comparison of the Proposed Framework with Selected Related Studies

## 4.4 Discussion of Findings

### 4.4.1 Contribution of Feature Fusion and Random Forest

Final Representation is composed of two distinct kinds of learned features. ConvNeXt-Tiny provides convolutional features characterizing local and hierarchical structure in the retina, and Swin-V2-Tiny contributes shifted-window attention to bring context. There is not a major improvement from fusion, yet concatenation of both embeddings contributed to a higher number of predictors for the Random Forest to consider. Then Random Forest could fit a non-linear relationship among those predictors without needing a new end-to-end training phase of deep neural networks. There is a noticeable gap between training accuracy and accuracy on held out images, and this means that the Random Forest fits well the training set embeddings. In order to alleviate this issue, the chosen model included small trees, large minimum splits and leaves size, restricted feature sampling, bootstrap sampling, and balanced subsample weighting. However, a gap of 14.04 between training and testing accuracy persisted, showing that overfitting is still an issue. On the other hand, validation and test sets accuracy are separated by 1.25 percentage points.

### 4.4.2 Interpretation in Relation to Related Work

The proposed multiclass accuracy at 84.91% is close to the 85.10% achieved by Youldash et al. [16] and higher than some other results attained in IDRiD and other Kaggle diabetic retinopathy projects [21], [26]. There are much higher numbers in other parts of the literature cited here, although the studies in question do not all employ the same database, number of classes, preprocessing step, or splitting

method. As such, the top line results cannot be compared directly with the current protocol.

One advantage of this experimental protocol is its separation of model development choices from final testing. Here, the same preprocessing and split procedure were maintained throughout, stochastic augmentation was done only during training, deep checkpoint selection was based on the validation step, and the parameters of Random Forest models were selected independently of test labels. Both aggregate and class-specific metrics make this weakness clear.

## 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.

It is computationally intensive as well since two deep backbones need to be fine-tuned and two embedding models need to be generated prior to training the classical classifier.

Various setups of Random Forests will be experimented with before the optimal model is chosen.

## 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 generated convolutional features for describing local and hierarchical retinal structure, while Swin-V2-Tiny offered an additional embedding that accounted for larger spatial relationships using shifted window attention. Each model extracted 768 features. The combination of these resulted in a 1,536-dimensional feature set, which was then classified using a Random Forest. By separating feature extraction from classification, it was possible to store the embeddings and use them for comparing various Random Forest setups. All in all, the experiment shows that a hybrid representation that combines convolution and transformer can work well when used with Random Forest for five-class diabetic retinopathy grading. ConvNeXt-Tiny and Swin-V2-Tiny extract different types of information in images, and combining them allows the classifier to utilize both. However, the current results should be considered as dataset-specific due to the fact that validation was done 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] 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.
- [15] Guefrachi, S., Ehtioui, A., & Hamam, H. (2025). Diabetic retinopathy detection using deep learning multistage training method. *Arabian Journal for Science and Engineering*, 50, 1079-1096.

- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] Alwakid, G., Gouda, W., & Humayun, M. (2023). Enhancement of diabetic retinopathy prognostication using deep learning, CLAHE, and ESRGAN. *Diagnostics*, 13(14), 2375.
- [21] Lin, C.-L., & Wu, K.-C. (2023). Development of revised ResNet-50 for diabetic retinopathy detection. *BMC Bioinformatics*, 24, 157.
- [22] 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.
- [23] 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.

- [24] Ali, A. A., & Dawood, F. A. A. (2023). Deep learning of diabetic retinopathy classification in fundus images. *Journal of Engineering*, 29(12), 139-152.
- [25] 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.
- [26] 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.
- [27] 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.
- [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.