

Faculty of Artificial intelligence Engineering

Department of AI & Data Science

**Early Detection of Diabetic Retinopathy**

**Using Deep Learning**

Senior 1 Project

Prepared by

Mohammad Barazi

Mohammad Khir Al homsi

Supervised by

Dr. Nisreen Sulayman

Eng. Zeinab Baghdadi

2025-2026

Faculty of Artificial intelligence Engineering

Department of AI & Data Science

**SUPERVISOR CERTIFICATION**

I certify that the preparation of this project **Early Detection of Diabetic Retinopathy Using Deep Learning**

prepared by **Mohammad Barazi & Mohammad Khir Alhomsi**

Was made under my supervision at the Faculty of Engineering in partial fulfillment of the Requirements for the Degree of Bachelor

**Name:**

**Date:**

**Signature:**

**Abstract**

This project presents an AI-based system for automated grading of diabetic retinopathy severity from retinal fundus images using deep learning. The goal is to develop a robust and scalable model capable of classifying diabetic retinopathy into five ordinal grades (0–4) on the APTOS 2019 dataset. The dataset contains 3,662 clinician-labeled images and exhibits a clear class imbalance across the five grades, which can bias standard classifiers toward the majority class. To address these challenges, a transfer-learning approach was adopted using an EfficientNetB3 backbone pretrained on ImageNet, with a regression-based output head to better reflect the ordinal nature of the labels. Key preprocessing steps included removing dark borders/background artifacts via image cropping, resizing to a fixed resolution, and normalization to reduce variability arising from different acquisition conditions. Training was conducted using stratified 5-fold cross-validation with class-weighted optimization to mitigate imbalance effects. In addition, an optimized thresholding strategy was applied to map continuous predictions to discrete grades, explicitly maximizing the Quadratic Weighted Kappa (QWK), the standard metric for APTOS 2019. The model achieved strong validation performance with a mean best-fold QWK of 0.860 (std 0.014) and a peak fold QWK of 0.887. The results indicate that combining ordinal-aware modeling with threshold optimization can improve agreement with clinical grading, supporting practical use cases such as large-scale screening and decision support, while highlighting opportunities for future work in external validation, interpretability, and more advanced imbalance handling.

**الملخص**

يقدّم هذا المشروع نظاماً قائماً على الذكاء الاصطناعي للتصنيف الآلي لشدّة اعتلال الشبكية السكري اعتماداً على صور قاع العين باستخدام تقنيات التعلّم العميق. يهدف العمل إلى تطوير نموذج قوي وقابل للتوسّع قادر على تصنيف اعتلال الشبكية السكري إلى خمس درجات مرتّبة (0–4) بالاعتماد على مجموعة بيانات APTOS 2019. تحتوي مجموعة البيانات على 3,662 صورة مُعنونة من قبل مختصين، وتُظهر اختلالاً واضحاً في توازن الفئات عبر الدرجات الخمس، ما قد يسبّب انحياز النماذج التقليدية نحو الفئة الأكبر. لمواجهة هذه التحديات، تم اعتماد منهجية نقل التعلّم باستخدام بنية EfficientNetB3 المُدرّبة مسبقاً على ImageNet، مع رأس إخراج بصيغة انحدار (Regression) لتمثيل الطبيعة الرتبية للتصنيفات بشكل أدق. شملت خطوات المعالجة المسبقة الأساسية إزالة الحواف/الخلفية الداكنة عبر قصّ الصورة، وتغيير الحجم إلى أبعاد ثابتة، وإجراء التطبيع لتقليل التباين الناتج عن اختلاف ظروف التصوير. تم تدريب النموذج باستخدام تحقق متقاطع Stratified K-Fold بخمس طيات مع أوزان للفئات للتخفيف من أثر عدم التوازن. إضافةً إلى ذلك، تم تطبيق استراتيجية تحسين العتبات (Optimized Thresholding) لتحويل التنبؤات المستمرة إلى درجات منفصلة مع تعظيم معيار **Quadratic Weighted Kappa (QWK)** بوصفه المقياس القياسي في APTOS 2019. حقق النموذج أداءً قوياً على بيانات التحقق بمتوسط أفضل QWK عبر الطيات قدره **0.860** (انحراف معياري **0.014**) وبقيمة قصوى وصلت إلى **0.887**. تشير النتائج إلى أن دمج النمذجة الحسّاسة للرتبية مع تحسين العتبات يمكن أن يرفع مستوى الاتفاق مع التقييم السريري، بما يدعم تطبيقات عملية مثل الفحص واسع النطاق ودعم القرار، مع إبراز فرص عمل مستقبلية تشمل التحقق الخارجي، ورفع قابلية التفسير، وتبنّي أساليب أكثر تقدماً لمعالجة عدم توازن الفئات.

Table of Contents

[Chapter 1 Introduction 13](#_Toc220522522)

[1.1.Introduction 14](#_Toc220522523)

[1.2.Problem Definition 14](#_Toc220522524)

[1.3.Project Objectives 15](#_Toc220522525)

[1.4. Tools and concepts 16](#_Toc220522526)

[1.4.1Using EfficientNetB3 for Transfer Learning 17](#_Toc220522527)

[1.*4.2* The Core Idea 18](#_Toc220522528)

[Chapter 2 Literature Review 20](#_Toc220522529)

[1.1. Introduction 20](#_Toc220522530)

[2.2.Used Tools 20](#_Toc220522531)

[2.2.1Apply of Parsifal in Our Project 21](#_Toc220522532)

[2.3.Summary of the Accepted Papers: 22](#_Toc220522533)

[2.2.1. Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy 22](#_Toc220522534)

[2.3.1.2.Contribution 23](#_Toc220522535)

[2.3.2. A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric 23](#_Toc220522536)

[2.3.3. Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy 25](#_Toc220522537)

[2.3.4. Innovative Approach for Diabetic Retinopathy Severity Classification: An AI-Powered Tool using CNN-Transformer Fusion 27](#_Toc220522538)

[2.3.5. Leveraging FastViT based knowledge distillation with EfficientNet-B0 for diabetic retinopathy severity classification 28](#_Toc220522539)

[2.3.5.2. Contribution: 30](#_Toc220522540)

[2.4.Comparation 31](#_Toc220522541)

[2.5.Pros and Cons 32](#_Toc220522542)

[2.6.Conclusion 33](#_Toc220522543)

[Chapter 3.Dataset collection 34](#_Toc220522544)

[3.1.Introduction 34](#_Toc220522545)

[3.2.APTOS 2019 Dataset 35](#_Toc220522546)

[3.2.1.Structure of the APTOS 2019 Dataset : 35](#_Toc220522547)

[3.2.2.Availability 35](#_Toc220522548)

[3.2.3Data preprocessing 36](#_Toc220522549)

[3.3.Dataset Splitting for Training and Validation 37](#_Toc220522550)

[Chapter 4. Implementation 39](#_Toc220522551)

[4.1.Introduction 40](#_Toc220522552)

[4.2Objective of the Project 40](#_Toc220522553)

[4.3Dataset Overview 40](#_Toc220522554)

[4.4Model Architecture 42](#_Toc220522555)

[4.4.1. Model 1 — EfficientNetB3 (Regression Output + OptimizedRounder) 42](#_Toc220522556)

[4.4.2. Model 2 — EfficientNetB3 + Ordinal CORAL (4 logits) + Constrained Cutoffs 43](#_Toc220522557)

[4.5 Training Details 44](#_Toc220522558)

[4.5.1. Model 1 Training Details (Stratified 5-Fold CV + Sample Weights) 44](#_Toc220522559)

[4.5.2. Model 2 Training Details (Stratified Split + Oversampling + Mixed Precision) 45](#_Toc220522560)

[4.6Results 46](#_Toc220522561)

[4.6.1. Model 1 Results (5-Fold OOF QWK + Training Curve) 46](#_Toc220522562)

[4.6.2. Model 1 Confusion Matrix (OOF Error Analysis) 47](#_Toc220522563)

[4.6.3. Model 2 Results (Best QWK + Training Curve) 48](#_Toc220522564)

[4.6.4. Model 2 Confusion Matrix (Best QWK) 49](#_Toc220522565)

[4.7. Comparison Between the Two Models (Results + Why Both Are Good) 50](#_Toc220522566)

[5.7.1. Comparison Summary 50](#_Toc220522567)

[4.7.2. Results Comparison (Important Note) 51](#_Toc220522568)

[4.7.3. Why Keeping Both Models Strengthens the Project 51](#_Toc220522569)

[Chapter 5.Research Gaps and Future Direction 52](#_Toc220522570)

[5.1 Introduction 53](#_Toc220522571)

[5.2 Feature Representation Limited to Static Global Patterns 53](#_Toc220522572)

[5.3 Class Imbalance Handling Without Broader Comparison 54](#_Toc220522573)

[5.4 Metric Alignment and Ordinal Modeling Limitations 54](#_Toc220522574)

[5.5 Generalization Under Domain Shift and Image Quality Variability 55](#_Toc220522575)

[Chapter 6. Conclusion 56](#_Toc220522576)

[Chapter 7. References 59](#_Toc220522577)

Table of Figure

[Figure 1 Sample of Dataset APTOS 2019 36](#_Toc220524816)

[Figure 2 Sample of our Dataset 37](#_Toc220524817)

[Figure 3 Code that confirms stratified splitting (Model 1) 38](#_Toc220524818)

[Figure 4 Code that confirms stratified splitting (Model 2) 39](#_Toc220524819)

[Figure 5 Code that confirms the same classes exist in train and validation (Model 2) 39](#_Toc220524820)

[Figure 6 Model 1 Validation 47](#_Toc220524821)

[Figure 7 Model 1 Confusion Matrix 48](#_Toc220524822)

[Figure 8 Model 2 Validation 49](#_Toc220524823)

[Figure 9 Model 2 Confusion Matrix 50](#_Toc220524824)

Tables of tables

[Table 1 Shortcuts 8](#_Toc220467161)

[Table 2 Comparation 29](#_Toc220467162)

[Table 3 Pros & Cons 30](#_Toc220467163)

[Table 4Project charter 34](#_Toc220467164)

[Table 5Statement of Work (SOW) 35](#_Toc220467165)

[Table 6.Preliminary Scope Statement 38](#_Toc220467166)

[Table 7 Risk management 40](#_Toc220467167)

Table of Shortcuts:

Table 1 Shortcuts

| **Term** | **Full form / Meaning** |
| --- | --- |
| **AI** | **Artificial Intelligence** |
| **DR** | **Diabetic Retinopathy** |
| **APTOS 2019** | **Asia Pacific Tele-Ophthalmology Society 2019 Blindness Detection (Kaggle dataset/competition)** |
| **QWK** | **Quadratic Weighted Kappa** |
| **CV** | **Cross-Validation** |
| **OOF** | **Out-Of-Fold predictions** |
| **K-Fold CV** | **K-Fold Cross-Validation** |
| **SKFold** | **Stratified K-Fold (StratifiedKFold)** |
| **SSS** | **Stratified Shuffle Split (StratifiedShuffleSplit)** |
| **Class Imbalance** | **Unequal number of samples across classes** |
| **Oversampling** | **Increasing sampling frequency of minority classes during training** |
| **Sample Weights** | **Per-sample weights to reduce class imbalance bias** |
| **ROI** | **Region Of Interest (focus area in the image)** |
| **Fundus Image** | **Retinal fundus photograph used as model input** |
| **Preprocessing** | **Steps applied before training (crop/resize/normalize, etc.)** |
| **Crop From Gray** | **Cropping method to remove dark background around the retina** |
| **Resize** | **Scaling images to a fixed input size (e.g., 300×300 or 512×512)** |
| **Data Augmentation** | **Random transformations to improve generalization** |
| **RandomZoom** | **Data augmentation: random zoom in/out** |
| **RandomContrast** | **Data augmentation: random contrast changes** |
| **CNN** | **Convolutional Neural Network** |
| **EfficientNet** | **CNN family using compound scaling (depth/width/resolution)** |
| **EfficientNetB3** | **Specific EfficientNet variant used as the backbone** |
| **Backbone** | **Pretrained feature extractor (base network)** |
| **Transfer Learning** | **Using pretrained weights (e.g., ImageNet) for a new task** |
| **ImageNet** | **Large-scale dataset commonly used for pretraining vision models** |
| **Head** | **Final layers added on top of the backbone for the task output** |
| **GAP** | **Global Average Pooling** |
| **Dropout** | **Regularization technique that randomly drops activations** |
| **BN** | **Batch Normalization** |
| **BN Freeze** | **Keeping BatchNorm layers non-trainable during fine-tuning** |
| **Warm-up** | **Initial training stage with frozen backbone** |
| **Fine-tuning** | **Training stage where some backbone layers are unfrozen** |
| **LR** | **Learning Rate** |
| **LR Scheduling** | **Changing learning rate during training based on validation behavior** |
| **ReduceLROnPlateau** | **Callback that reduces LR when validation metric stops improving** |
| **EarlyStopping** | **Callback that stops training if no improvement is observed** |
| **ModelCheckpoint** | **Callback that saves the best model/weights during training** |
| **Adam** | **Adaptive Moment Estimation optimizer** |
| **AdamW** | **Adam optimizer with decoupled Weight Decay** |
| **WD** | **Weight Decay (regularization in optimizers like AdamW)** |
| **Huber Loss** | **Robust regression loss (less sensitive to outliers than MSE)** |
| **BCE** | **Binary Cross-Entropy loss** |
| **Logits** | **Raw model outputs before applying sigmoid/softmax** |
| **Sigmoid** | **Activation function mapping values to (0, 1)** |
| **Softmax** | **Activation function mapping logits to class probabilities (multi-class)** |
| **Regression Output** | **Predicting a continuous severity score (e.g., 0–4)** |
| **Naive Rounding** | **Converting continuous predictions to classes using round + clip** |
| **Threshold Optimization** | **Learning cutoffs to map scores → grades to maximize QWK** |
| **OptimizedRounder** | **Threshold optimizer (searches best cut points for QWK)** |
| **Cutoffs** | **Threshold values used to convert continuous/ordinal score into classes** |
| **Ordinal Labels** | **Labels with a natural order (0 < 1 < 2 < 3 < 4)** |
| **Ordinal Regression** | **Learning approach designed for ordered classes** |
| **CORAL** | **COnsistent RAnk Logits (ordinal regression framework)** |
| **CORAL Head** | **4-logit output layer representing 5 ordinal grades (0–4)** |
| **Mixed Precision** | **Training with float16/float32 to speed up and save memory** |
| **FP16** | **16-bit floating point precision** |
| **TF** | **TensorFlow** |
| **Keras** | **High-level API (used with TensorFlow)** |
| **tf.data** | **TensorFlow input pipeline API** |
| **AUTOTUNE** | **TensorFlow option for automatic pipeline performance tuning** |
| **Confusion Matrix** | **Table showing true vs predicted classes** |
| **Normalized Confusion Matrix** | **Confusion matrix scaled by row totals (per-true-class rates)** |
| **Precision** | **TP / (TP + FP)** |
| **Recall** | **TP / (TP + FN)** |
| **F1-score** | **Harmonic mean of precision and recall** |
| **Macro Avg** | **Average metric across classes (treats classes equally)** |
| **Weighted Avg** | **Average metric weighted by class support (frequency)** |
| **Accuracy** | **Correct predictions / total predictions** |

# Chapter 1 Introduction

## 1.1.Introduction

In this chapter, we introduce our project and explain the main motivations for building an automated system for **diabetic retinopathy (DR) severity grading**. We also outline the key challenges in the **APTOS 2019** retinal fundus dataset—most notably **class imbalance across the five grades (0–4)** and variability in image quality—and why these issues require a carefully designed deep learning pipeline. Finally, we present the objectives we aim to achieve and provide an overview of the main concepts and tools used in our model development and evaluation.

## 1.2.Problem Definition

Diabetic retinopathy (DR) is a diabetes-related eye disease that occurs when chronically high blood sugar damages the small blood vessels in the retina, leading to leakage, bleeding, retinal ischemia, and in advanced cases abnormal new vessel growth that can threaten vision. DR can progress silently for a long time, and if it is not detected and treated early, it may result in irreversible vision loss or preventable blindness.

The clinical challenge is not only detecting whether DR exists, but accurately grading its **severity stages**. DR severity is typically described along a spectrum from no DR to increasing non-proliferative stages and finally proliferative DR, and each stage is associated with specific retinal lesions (e.g., microaneurysms, hemorrhages, venous beading, IRMA, and neovascularization). These lesion patterns can be subtle, small, and sometimes difficult to differentiate—especially when disease changes are mild or when two adjacent grades share similar visual signs. As a result, distinguishing between neighboring grades (for example mild vs moderate, or severe vs proliferative) can be clinically challenging, even for experienced graders.

In real-world practice, manual grading depends on expert interpretation of fundus photographs, which introduces several difficulties. First, fundus image quality can vary (illumination, focus, blur, artifacts, and partial field of view), which can hide key lesions and make the grading decision uncertain. Second, DR grading contains a subjective component, and multiple studies report **inter-observer variability** between specialists when grading the same retinal photographs, meaning that two experts may not always assign the same severity grade to the same image. This variability can lead to inconsistent decisions regarding follow-up urgency and treatment pathways.

Another critical issue is access to eye care and screening capacity. Effective DR prevention requires large-scale periodic screening of people with diabetes, but many regions face shortages in trained eye-care professionals and unequal access to specialized services. This means that not all patients can be screened in time, and even when screening is available, high workload and limited resources can increase the chance of delayed review or missed/incorrect grading in borderline cases. Reports discussing DR screening barriers highlight factors such as shortage of ophthalmologists, geographic remoteness, and limited availability of appropriate diagnostic equipment, which collectively reduce coverage and early detection.

Therefore, the motivation of this project is to support DR screening and grading by developing an AI-based model that can analyze retinal fundus images and provide consistent severity predictions. By automating grading, the system can help reduce dependence on limited specialist availability, improve scalability for large diabetic populations, and assist clinicians by offering a second opinion—especially in cases where visual differences between grades are subtle and easy to overlook.

## 1.3.Project Objectives

The objectives of a diabetic retinopathy (DR) grading project focus on enabling **early and reliable screening** by automatically assessing retinal fundus images and assigning a **five-grade severity level (0–4)** in alignment with the APTOS 2019 labeling scheme. A primary goal is to support clinical decision-making by helping identify patients at higher risk so that **timely referral and treatment** can be provided, since early detection and timely management are critical to avoiding irreversible vision loss. Another objective is to build a model that remains dependable under real-world data challenges, particularly the **class imbalance** present in APTOS 2019 and the variability in image acquisition conditions, which can cause models to over-predict the majority class if not properly addressed. In addition, because DR grades are **ordinal**, the project aims to evaluate and optimize the system using metrics that reflect clinical agreement—most importantly **Quadratic Weighted Kappa (QWK)**, which penalizes larger grade disagreements more than minor errors and is widely used for DR grading in this context. Ultimately, the project seeks to deliver a robust, scalable deep learning pipeline that can contribute to more efficient DR screening workflows, while providing a clear basis for future improvements such as stronger generalization, enhanced interpretability, and more advanced handling of imbalanced and heterogeneous medical image data.

## 1.4. Tools and concepts

EfficientNetB3:

EfficientNetB3 is a convolutional neural network from the EfficientNet family introduced in the 2019 paper **“EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”**. It is designed to achieve high accuracy with improved computational efficiency, and it is well-suited for transfer learning in medical image classification tasks such as **five-grade diabetic retinopathy (0–4) grading**.   
• **Architecture Design:** EfficientNet models are built from an optimized baseline (EfficientNet-B0) and scaled into variants (including B3) using a consistent scaling strategy:  
1 baseline network designed via NAS for efficiency (B0)  
2 scaled variants (B1–B7) derived from the baseline  
3 EfficientNetB3 as one of these compound-scaled variants  
• **Compound Scaling:** Scales network depth, width, and input resolution together using a single compound coefficient instead of scaling one dimension independently.  
• **Core Building Blocks:** Uses **MBConv (mobile inverted bottleneck)** blocks and adds **squeeze-and-excitation (SE)** to improve representational power while keeping computation manageable.   
• **Pre-trained Availability:** Commonly available with **ImageNet pretrained weights**, enabling transfer learning and faster convergence when training on APTOS 2019.

### 1.4.1Using EfficientNetB3 for Transfer Learning

There are two main strategies for transfer learning using EfficientNetB3:

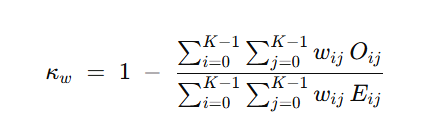
1. **Feature Extraction (Freeze Base, Replace Head):**  
   Purpose: Use EfficientNetB3’s convolutional base to extract features from fundus images, then train a new classification/regression head on top.  
   Use Cases:  
   When your dataset is limited in size and you want faster training with lower risk of overfitting.
2. **Fine-Tuning (Unfreeze Some Layers):**  
   Purpose: Allow the network to adapt more deeply to APTOS 2019 by unfreezing and updating the weights of some layers (often the higher layers) after an initial warm-up.  
   Use Cases:  
   When the dataset is more domain-shifted from ImageNet (as in medical images) and additional compute/tuning is available for improved performance.

**Quadratic Weighted Kappa (QWK)**  
Quadratic Weighted Kappa is an agreement-based metric commonly used for **ordinal classification** problems where labels have a natural order (such as diabetic retinopathy grades 0–4). It measures agreement between the clinician’s grade and the predicted grade while penalizing larger disagreements more heavily than small ones. In the **APTOS 2019** competition setting, submissions are evaluated using QWK.

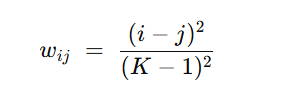
### 1.*4.2* The Core Idea

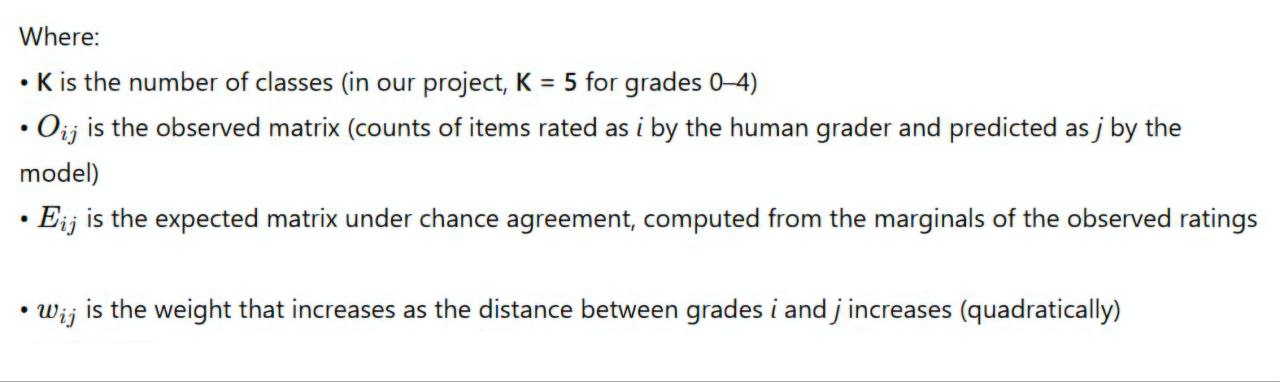
Traditional metrics such as accuracy treat all mistakes equally, which does not reflect the clinical meaning of severity grading. QWK explicitly accounts for the **ordinal distance** between grades, so predicting a far-away grade (e.g., 0 instead of 4) is penalized more than a near miss (e.g., 2 instead of 3). This aligns evaluation more closely with real-world DR severity assessment and is therefore the standard metric used for APTOS 2019 scoring.

#### Formula



For **quadratic** weighting, the disagreement weights are typically defined as:





# Chapter 2 Literature Review

## Introduction

In this chapter, we will review the literatures summary them and we will mention the gaps of each one, additionally we will compare between them.  
In our project, which focuses on **Diabetic Retinopathy (DR) severity grading from fundus images using deep learning** on **APTOS 2019 Blindness Detection** dataset, we used this chapter to collect the most related studies to our topic.

## 2.2. Used Tools

## 

To collect and organize the related works for our topic, we used academic search sources that are suitable for medical AI and deep learning papers such as:

* **PubMed / PMC** (medical and clinical papers)
* **IEEE Xplore** and similar digital libraries (engineering & AI papers)
* **arXiv** (recent deep learning methods like EfficientNet و CORAL)

Also, we used the official competition description to confirm the dataset problem setting and the evaluation metric, where submissions are scored using **Quadratic Weighted Kappa (QWK)**.

### 2.2.1 Apply of Literature Searching in Our Project

We began by defining our research questions

Use advanced search with keywords like:

* APTOS 2019 Blindness Detection diabetic retinopathy grading
* Quadratic Weighted Kappa (QWK) diabetic retinopathy APTOS
* EfficientNet diabetic retinopathy fundus images
* Ordinal regression CORAL severity grading

Study Selection is a crucial step in conducting a Systematic Literature Review, this step follows your “review protocol”, especially the “inclusion and exclusion criteria” you defined earlier. Sush as:

* Include: papers about **DR grading (0–4)** using **fundus images** and deep learning, and papers that use **QWK** as evaluation (because labels are ordinal).
* Exclude: papers not related to fundus DR grading, or papers without clear evaluation.

## 2.3.Summary of the Accepted Papers:

### 2.2.1. Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy

Cheena Mohanty, · Sakuntala Mahapatra · Biswaranjan Acharya · et al.  
Published online: 19 June 2023

#### 2.3.1.1.Paper summary:

**Dataset:**

The study evaluated the proposed models on the **APTOS 2019 Blindness Detection** dataset.

**Problem:**

The paper highlights that the dataset has an **imbalanced class distribution**, which affects classification behavior and model bias.

**Preprocessing:**

The authors report applying preprocessing to the retinal fundus images before training the models.

**Models compared:**

* A **hybrid architecture** that combines **VGG16** (feature extraction) with an **XGBoost classifier**.
* A **DenseNet121** model for DR classification.

**Evaluation:**

The comparison was mainly reported using **accuracy** as the evaluation metric.

**Results reported:**

DenseNet121 achieved higher performance than the hybrid model on APTOS 2019.

## 2.3.1.2.Contribution

Algorithm: Two DL solutions for DR grading on APTOS 2019, including a hybrid **VGG16 + XGBoost** pipeline and an end-to-end **DenseNet121** model.   
Parameters: The paper emphasizes applying preprocessing and using balancing techniques due to the imbalance of APTOS 2019 classes.   
Results Achieved:

* Hybrid model accuracy: **79.50%**
* DenseNet121 accuracy: **97.30%** (reported as superior on the same dataset).

### 2.3.2. A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric

Sai Venkatesh Chilukoti, · Liqun Shan · Vijay Srinivas Tida · et al.  
Published online: 06 February 2024

#### 2.3.2.1Paper summary:

**Main idea:**

The paper targets DR grading and argues that metrics like accuracy/F1 can be misleading in DR grading because the classes are **ordered** and the severity of misclassification matters.

**Metric focus:**

The authors prioritize **Quadratic Weighted Kappa (QWK)** as the main grading metric.

**Preprocessing / Enhancement:**

They mention using **CLAHE** for enhancement and **Gaussian blur** for noise reduction (along with normalization in different strategies).

**Backbone & classifier:**

They use **EfficientNet-B3** as the main backbone and replace the classifier head with a designed **3-layer classifier** using FC layers + **dropout** + **ReLU**.

**Classifier details (as described):**

* EfficientNet-B3 feature output is pooled then flattened to **1536** features.
* FC(1536→512) → Dropout(0.5) → ReLU → FC(512→512) → Dropout(0.25) → ReLU → FC(→5 classes).

**Ensemble strategy:**

A key contribution is an ensemble method based on **combining predictions from saved weights** during training (example: saving at epochs 30 and 60 and aggregating).

**Training / image size detail:**

One strategy described training EfficientNet-B3 for **60 epochs**, and resizing images to **150×150**.

#### 2.3.2.2.Contribution:

Algorithm: Transfer learning + resource-efficient ensemble (weights-at-epochs ensemble) with EfficientNet-B3 + custom 3-layer classifier head for 5-class DR grading.   
Parameters:

* Image resize mentioned: **150×150**
* Training epochs example: **60**
* Dropout rates: **0.5** and **0.25** in the classifier head.   
  Results Achieved (reported QWK):
* **EyePACS:** 0.901
* **APTOS:** **0.967**
* **Messidor:** 0.944

### 2.3.3. Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy

Hossein Shakibania, · Sina Raoufi · Behnam Pourafkham · et al.  
Published online: 19 August 2023 (arXiv preprint) / 2024 journal version referenced

#### 2.3.3.1.Summary

 **Goal:**

The paper proposes a model for both DR detection and **5-stage grading (0–4)** using a single fundus image, and evaluates on **APTOS 2019**.

 **Architecture:**

A **dual-branch transfer learning** design using **EfficientNetB0** and **ResNet50** as feature extractors (two branches).

 **Data strategy:**

The authors train using a **merged multi-source dataset** and also describe a “selectively merged” dataset strategy, then evaluate on APTOS 2019.

 **Class imbalance handling:**

They report using **augmentations** and show comparisons of original vs augmented class distributions.

 **Regularization:**

Dropout is used (reported as **0.25**) after each linear layer in the fully connected network to reduce overfitting.

 **Hyperparameter tuning:**

They report Bayesian optimization (Ray Tune) and provide tuned values for learning rate and momentum.

 **Evaluation:**

They report multiple metrics including **QWK, accuracy, precision, sensitivity, specificity, F1, and AUC** for multi-class grading.

 **Limitation mentioned:**

The paper notes that severe and proliferative stages can be difficult to separate due to visual similarity and limited samples.

#### 2.3.3.2.Contribution:

Algorithm Used: Dual-branch transfer learning model combining **EfficientNetB0 + ResNet50** feature representations for DR grading.   
Parameters:

* Tuned hyperparameters: learning rate **0.06**, momentum **0.66** (reported from Bayesian optimization).

Dropout: **0.25**  in FC layers.   
Results Achieved (APTOS 2019 stage grading, reported):

* **QWK:** **93.00%**
* **Accuracy:** **89.60%**
* **F1-score:** 89.15%
* **Specificity:** 97.72%

### 2.3.4. Innovative Approach for Diabetic Retinopathy Severity Classification: An AI-Powered Tool using CNN-Transformer Fusion

Khosro Rezaee, · Fateme Farnami  
Published online: 01 April 2025

#### 2.3.4.1.Summary

 **Goal:**

Improve DR detection and severity classification by combining **CNN** (local feature extraction) with **Transformer** (long-range dependencies).

 **Dataset:**

The model is trained and evaluated on **APTOS 2019**, with additional validation on **IDRiD** to test generalization.

 **Imbalance handling:**

The authors state that training was performed on an **augmented APTOS 2019** dataset to address class imbalance.

 **Evaluation metrics:**

The paper reports accuracy and also mentions AUC, specificity, and sensitivity as evaluation metrics.

 **Main outcome:**

The reported results indicate strong performance on APTOS and consistent accuracy on IDRiD.

#### 2.3.1.2.Contribution:

Algorithm Used: CNN–Transformer fusion for DR classification and severity assessment trained with augmentation to handle imbalance.

Parameters: The paper emphasizes augmentation-based handling of imbalance and evaluation across datasets (APTOS + IDRiD).

Results Achieved (reported):

* Accuracy on **APTOS 2019:** **94.28%**
* Accuracy on **IDRiD:** **95.23%**

### 2.3.5. Leveraging FastViT based knowledge distillation with EfficientNet-B0 for diabetic retinopathy severity classification

Jyotirmayee Rautaray, · Ali B.M. Ali · Meenakshi Kandpal · et al.  
Published online: 28 June 2025 (Epub) / August 2025 issue

#### 2.3.5.1.Summary

**Main idea:** The study proposes **FastEffNet**, which uses transformer-based **knowledge distillation (KD)** to build a lightweight but accurate DR severity classifier.

**Teacher–Student setup:**

* Teacher model: **FastViT-MA26**
* Student model: **EfficientNet-B0**

**Dataset:**

The work uses APTOS blindness detection dataset with **3662 images** and **5 severity classes**.

**Preprocessing & imbalance:**

They mention preprocessing, normalization, splitting, and augmentation specifically to address **class imbalance**.

**Comparisons:**

The study compares additional student backbones such as **HGNet, ResNet50, MobileNetV3, and DeiT**.

**Interpretability:**

They apply **Grad-CAM++** to highlight important retinal regions influencing decisions.

**Evaluation:**

The paper reports multiple metrics (accuracy, precision, recall, F1, Cohen’s kappa, weighted kappa, MCC, AUC) to assess robustness.

### 2.3.5.2. Contribution:

Algorithm Used: Transformer-teacher distillation (FastViT) into a lightweight CNN student (EfficientNet-B0) for 5-class DR severity grading on APTOS.   
Parameters:

* Teacher: **FastViT-MA26**
* Student: **EfficientNet-B0**
* Computational cost (student): **0.38 GFLOPs** (reported).   
  Results Achieved (best student reported – EfficientNet-B0):
* Accuracy: **95.39%**
* Precision: **95.43%**
* Recall: **95.39%**
* F1-score: **95.37%**
* CKS: **0.94**, WKS: **0.97**, MCC: **0.94**
* AUC: **0.99**

## 2.4.Comparation

Table 2 Comparation

| **Paper Title** | **Dataset Used** | **Training Approach** | **AI Model** | **Evaluation & Result** |
| --- | --- | --- | --- | --- |
| Mohanty et al. (2023) | APTOS 2019 | DL classification + balancing | DenseNet121 / Hybrid | Acc: 97.30% (DenseNet121) ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC10301863/)) |
| Chilukoti et al. (2024) | APTOS 2019 + others | Transfer learning + ensemble | EfficientNet-B3 | QWK (APTOS): 0.967 ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC11323616/)) |
| Shakibania et al. (2023/2024) | APTOS 2019 + multi-center | Dual-branch transfer learning | Two pretrained backbones | QWK: 93.00, Acc: 89.60% ([arxiv.org](https://arxiv.org/abs/2308.09945)) |
| Rezaee & Farnami (2025) | APTOS 2019 | CNN + Transformer fusion | Hybrid CNN-Transformer | Acc: 94.28% ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC12009469/)) |
| Rautaray et al. (2025) | APTOS 2019 | Knowledge distillation | FastViT → EfficientNet-B0 | Acc: 95.39%, WKS: 0.97 ([sciencedirect.com](https://www.sciencedirect.com/science/article/pii/S2472630325000834)) |

5

## 2.5.Pros and Cons

Table 3 Pros & Cons

| **Paper Title** | **Pros** | **Cons** |
| --- | --- | --- |
| **Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy (2023)** | - Uses **APTOS 2019** and explicitly discusses the **class imbalance** challenge.- Provides a **comparative study** between two different approaches (end-to-end DL vs hybrid pipeline).- Includes a clear pipeline idea: preprocessing + training + evaluation, which is useful as a baseline reference for DR grading projects. ([mdpi.com](https://www.mdpi.com/1424-8220/23/12/5726?utm_source=chatgpt.com)) | - Focuses mainly on **accuracy** reporting, which may be less suitable for **ordinal DR grading** compared to kappa-based metrics.- The hybrid approach (CNN features + XGBoost) adds extra steps and can reduce simplicity/reproducibility compared to an end-to-end model. ([mdpi.com](https://www.mdpi.com/1424-8220/23/12/5726?utm_source=chatgpt.com)) |
| **A reliable diabetic retinopathy grading via transfer learning and ensemble learning with QWK metric (2024)** | - Strong point: prioritizes **Quadratic Weighted Kappa (QWK)** as a main metric for grading, which matches the **ordered severity labels**.- Uses **transfer learning (EfficientNet)** and proposes a **computationally efficient ensemble** idea (leveraging saved model states/weights).- Mentions using **augmentation** to improve generalization and handle limited data. ([SpringerLink](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02446-x?utm_source=chatgpt.com)) | - Ensemble methods can increase **inference complexity** (multiple models/weights or aggregation step).- Transfer learning performance can be sensitive to **preprocessing choices** and dataset-specific tuning, so results may vary across implementations. ([SpringerLink](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02446-x?utm_source=chatgpt.com)) |
| **Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy (2023/2024)** | - Proposes **dual-branch feature extraction** (two pretrained backbones), which can capture complementary features for DR grading.- Evaluates on **APTOS 2019** and reports results for both **detection** and **stage grading**, giving a more complete view of performance.- Reports **QWK** and other metrics, which supports ordinal-grade evaluation. ([arXiv](https://arxiv.org/abs/2308.09945?utm_source=chatgpt.com" \o "Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy)) | - Dual-branch architecture is typically **heavier computationally** (training + inference) than single-backbone models.- Combining/mixing datasets and heavy augmentation can introduce **domain shift** concerns if not carefully controlled. ([arXiv](https://arxiv.org/abs/2308.09945?utm_source=chatgpt.com" \o "Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy)) |
| **Innovative Approach for DR Severity Classification: CNN-Transformer Fusion (2025)** | - Combines **CNN (local features)** with **Transformer (long-range dependencies)** which is suitable for complex fundus patterns.- Uses **augmented APTOS 2019** and states it addresses **class imbalance** with augmentation.- Adds external validation on another dataset (IDRiD) to discuss **generalization**. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC12009469/?utm_source=chatgpt.com)) | - CNN-Transformer fusion models are usually **more complex** and can require more compute/memory than plain CNN transfer learning.- Performance can depend strongly on augmentation strategy; if augmentation is aggressive, it may risk learning non-clinical artifacts. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC12009469/?utm_source=chatgpt.com)) |
| **Leveraging FastViT knowledge distillation with EfficientNet-B0 for DR severity classification (2025)** | - Uses **knowledge distillation** (teacher → student) to get strong performance while keeping the student model **lightweight** for deployment.- Uses APTOS and reports multiple metrics (including **kappa/weighted kappa**) for grading evaluation.- Focuses on reducing computation while maintaining accuracy, which supports practical clinical/edge use. ([ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2472630325000834?utm_source=chatgpt.com)) | - Distillation adds an extra training stage (teacher training + distillation), which increases pipeline complexity.- Results depend on the **teacher quality** and distillation settings; reproducing the same gains may require careful tuning. ([ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2472630325000834?utm_source=chatgpt.com)) |

## 2.6.Conclusion

The most effective approaches for APTOS 2019 DR grading in recent studies (2023 and above) are based on deep learning with transfer learning backbones and modern improvements such as **ensembles**, **CNN-Transformer fusion**, and **knowledge distillation**, while focusing on evaluation metrics suitable for ordinal labels like **Quadratic Weighted Kappa (QWK)**.

# Chapter 3. Dataset collection

## 3.1. Introduction

In this chapter, we will discuss the dataset used in our project. Datasets play a crucial role in the success of any trained model, because they provide the information that the model learns from and uses to make predictions.

## 3.2. APTOS 2019 Dataset

The **APTOS 2019 Blindness Detection** dataset is designed to support research on automated **diabetic retinopathy (DR)** grading from retinal fundus images. Each image is labeled by a clinician with a severity grade on a **five-level scale (0–4)**, where 0 indicates *No DR* and 4 indicates *Proliferative DR*.

The dataset provides **3,662 labeled training images**, and a separate **test set of 1,928 images** (without public labels in the original competition setting).

### 3.2.1. Structure of the APTOS 2019 Dataset:

The dataset is typically organized as follows:

* **Fundus Images:** Each sample is a retinal fundus photograph stored as an image file. Image quality varies due to different acquisition conditions (illumination, focus, borders, and background).
* **Labels (Diagnosis):** Each training image has a ground-truth label representing DR severity from **0 to 4**.
* **Metadata File:** A CSV file (commonly train.csv) links each **image id** to its **diagnosis label (0–4)**.

### 3.2.2. Availability

The APTOS 2019 dataset is publicly available through **Kaggle** as part of the APTOS 2019 Blindness Detection competition resources and is widely used for research and benchmarking in DR severity classification.   
A key characteristic of this dataset is the **class imbalance** across the five grades. A commonly reported distribution for the training set is:

* Grade 0: 1805
* Grade 1: 370
* Grade 2: 999
* Grade 3: 193
* Grade 4: 295

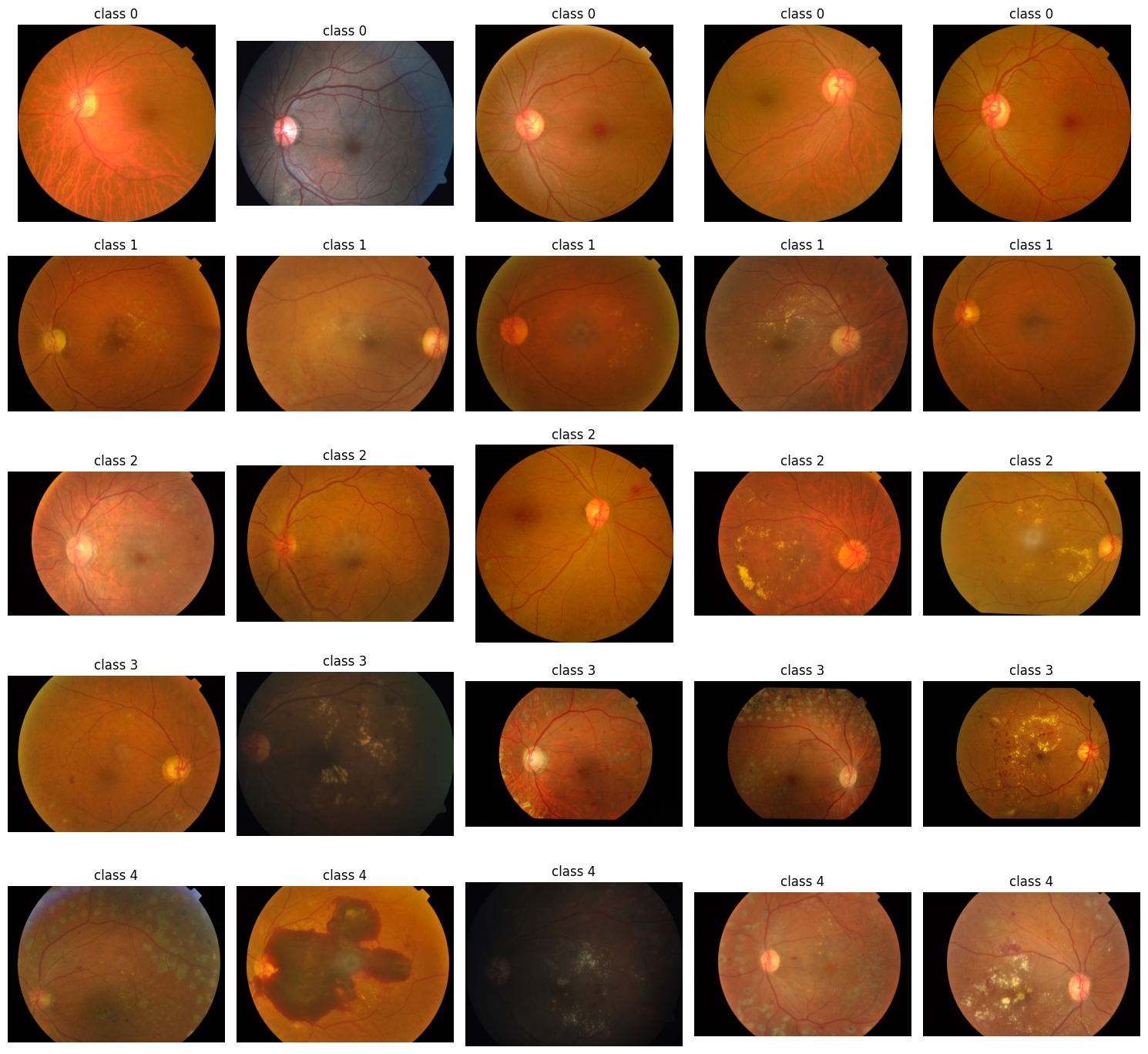


Figure 1 Sample of Dataset APTOS 2019

### 3.2.3Data preprocessing

To improve training stability and reduce noise caused by real-world acquisition variability, we applied preprocessing steps tailored for fundus images:

* **Border/Background Removal (Cropping):** Many fundus images include dark borders or non-informative background regions. Removing these regions helps the model focus on the retinal area. This is a common preprocessing step in APTOS-based pipelines.
* **Resizing and Standardization:** All images were resized to a fixed input size required by the model to ensure consistent tensor shapes during training.
* **Data Augmentation:** We used augmentation (e.g., flips/rotations/zoom/contrast) to increase variability and reduce overfitting, which is especially helpful when some grades have fewer samples.

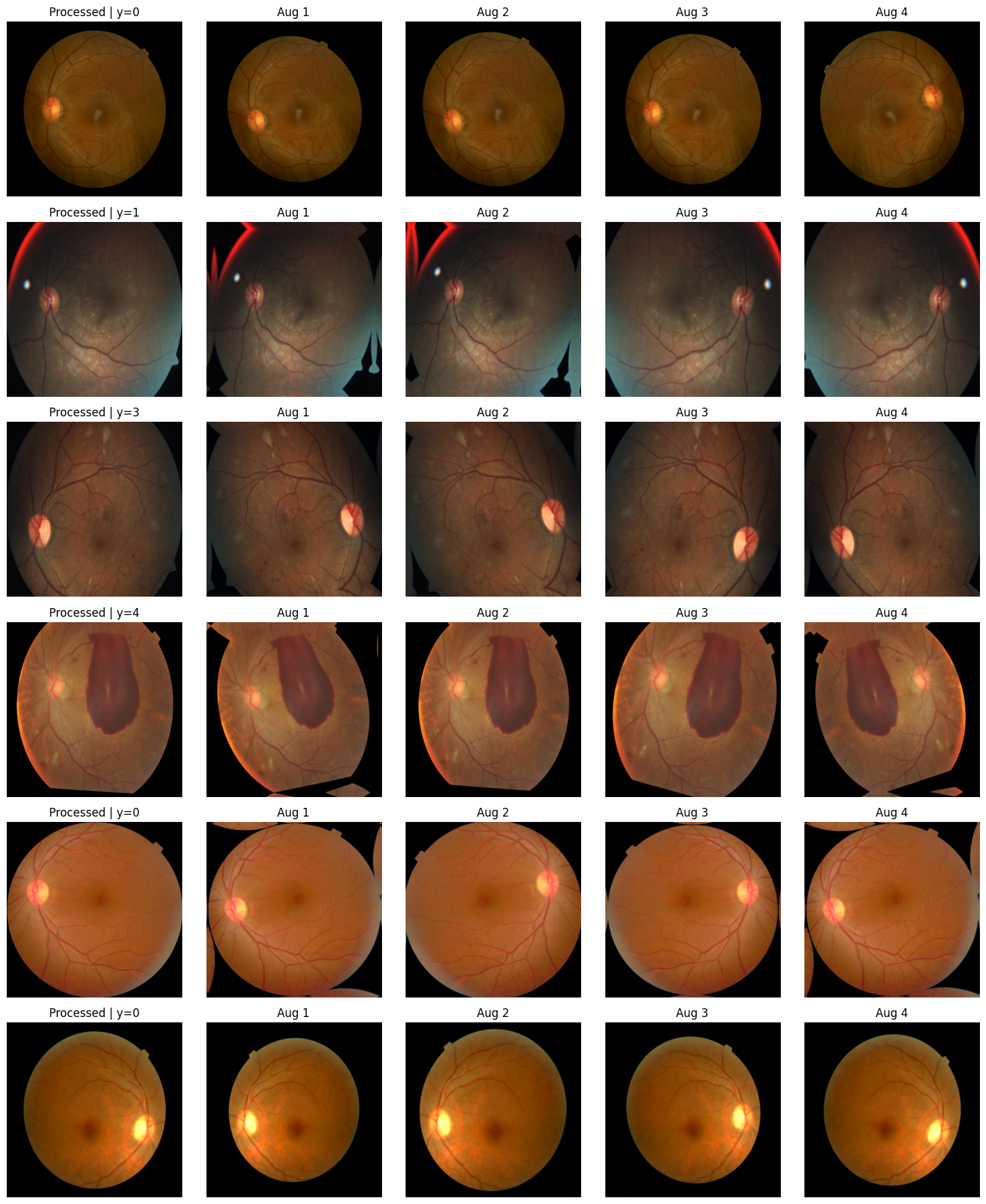


Figure Sample of our Dataset

## 3.3. Dataset Splitting for Training and Validation

Because the dataset is imbalanced, we used a validation strategy that preserves grade distribution across splits to produce more reliable evaluation results. This helps ensure that minority grades appear in both training and validation subsets and reduces evaluation bias caused by random splitting.

**Model 1 (Regression + 5-Fold CV) — Split Ratio**

* We used **Stratified 5-Fold Cross-Validation** (n\_splits = 5).
* This means that in **each fold**:
  + **Training ≈ 80%** of the data
  + **Validation ≈ 20%** of the data
* Stratification ensures that each fold preserves the percentage of samples for each class.

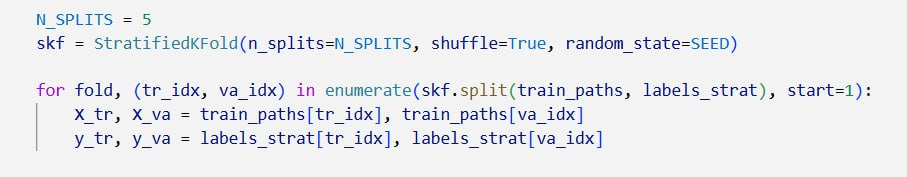


Figure Code that confirms stratified splitting (Model 1)

(This code uses labels\_strat in skf.split(...), which is exactly what makes the split stratified.)

**Model 2 (Ordinal CORAL) — Split Ratio**

* We used **Stratified 80/20 split** using StratifiedShuffleSplit with test\_size = 0.20.
* In your run, it produced:
  + **Train: 2929 images (~80%)**
  + **Validation: 733 images (~20%)**
* StratifiedShuffleSplit preserves class proportions across train and validation.

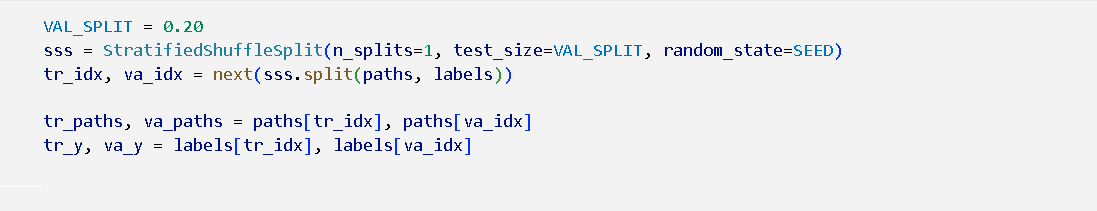


Figure Code that confirms stratified splitting (Model 2)



Figure Code that confirms the same classes exist in train and validation (Model 2)

**The printed distributions:**

* Train dist: [1444, 296, 799, 154, 236]
* Val dist: [361, 74, 200, 39, 59]  
  This confirms that **all five grades (0–4)** exist in both training and validation sets.

# Chapter 4. Implementation

## 4.1. Introduction

In this chapter, we will talk about our models, how the training goes, fine-tuning details, and all the evaluation metrix used in our project. Our project is based on **APTOS 2019 Blindness Detection**, where the official evaluation metric is **Quadratic Weighted Kappa (QWK)** for the 5 ordered grades (0–4).

4.2Objective of the Project

The goal of this project is to build an artificial intelligence system capable of classifying retinal fundus images into **five diabetic retinopathy severity grades (0–4)** using the **APTOS 2019** dataset. The system learns discriminative visual patterns from the retinal images and produces predictions that reflect the **ordinal nature** of DR severity, while being evaluated using QWK as in the APTOS 2019 setting.

4.3Dataset Overview

* **Dataset Name:** APTOS 2019 Blindness Detection
* **Task:** Five-grade DR severity classification **(0–4)**
* **Training images:** **3,662** labeled images
* **Test images:** **1,928** images (labels not provided in the original competition setting)
* **Main challenge:** Class imbalance and the ordinal nature of labels (0–4).

**Preprocessing and input preparation were implemented differently in the two models to test two strong pipelines:**

**Model 1 (300×300 pipeline):**

* ROI cropping using crop\_image\_from\_gray(tol=7) to remove dark borders/background.
* Resize to **300×300**.
* Minimal preprocessing (no heavy enhancement, only crop + resize + clipping).
* Augmentation:
  + In-model: RandomFlip / RandomRotation / RandomZoom / RandomContrast
  + In tf.data: random brightness + random saturation

**Model 2 (512×512 pipeline):**

* ROI crop SAFE to remove black background using mask-based crop **with safety checks** (avoid over-cropping and noisy cropping).
* Resize to **512×512**.
* Mixed precision enabled to handle high resolution efficiently.
* Augmentation (in-model): RandomFlip / RandomRotation / RandomZoom / RandomContrast

4.4Model Architecture

### 4.4.1. Model 1 — EfficientNetB3 (Regression Output + OptimizedRounder)

* **Backbone Model:** EfficientNetB3 (transfer learning, ImageNet weights). EfficientNet uses compound scaling (depth/width/resolution) to achieve strong accuracy–efficiency trade-off.
* **Input:** RGB fundus image **300×300**.
* **Head design:**
  + GlobalAveragePooling
  + Dropout(0.3)
  + Dense(1, sigmoid) then scaled to **0..4** using output = sigmoid(x) \* 4
* **Prediction type:** continuous score (0..4), then mapped to discrete classes (0–4).
* **Why this design:** regression-style outputs often work well with QWK because QWK penalizes larger mistakes more, and continuous outputs allow better tuning of boundaries using thresholds.

**Threshold mapping used in Model 1:**

* Baseline: naive rounding (round + clip).
* Improved: **OptimizedRounder** which searches thresholds to maximize QWK on OOF predictions.

### 4.4.2. Model 2 — EfficientNetB3 + Ordinal CORAL (4 logits) + Constrained Cutoffs

* **Backbone Model:** EfficientNetB3 (transfer learning).
* **Input:** RGB fundus image **512×512**.
* **Head design:**
  + GlobalAveragePooling
  + Dropout(0.40)
  + Dense(4, linear) → **4 logits** (CORAL encoding for 5 ordinal classes).
* **Ordinal encoding:** label (0–4) is converted into 4 ordered binary targets (y > 0, y > 1, y > 2, y > 3).
* **Loss:** BinaryCrossentropy(from\_logits=True) across the 4 ordinal outputs.
* **Why CORAL:** CORAL (Consistent Rank Logits) is designed for ordinal labels and reduces inconsistent ordinal predictions by enforcing rank-consistency.

**Score-to-grade mapping used in Model 2:**

* Logits → probabilities (sigmoid) → continuous severity score = sum(probabilities).
* Then a **constrained cutoffs search** is applied to find thresholds that maximize QWK while keeping predicted distribution realistic.

4.5 Training Details

### 4.5.1. Model 1 Training Details (Stratified 5-Fold CV + Sample Weights)

* **Cross-validation:** StratifiedKFold, **5 folds** (preserve class distribution in each fold).
* **Batch size:** 8
* **Loss:** Huber(delta=1.0)
* **Optimizer:** Adam
* **Warm-up stage:**
  + EfficientNetB3 frozen
  + Epochs = 2
  + LR = 1e-3
* **Fine-tuning stage:**
  + Backbone trainable, but freeze first 80% layers (train last 20%)
  + Freeze BatchNormalization layers during fine-tuning
  + Epochs = 10
  + LR = 1e-4
* **Callbacks:** ModelCheckpoint (monitor val\_qwk), EarlyStopping (monitor val\_qwk), ReduceLROnPlateau (monitor val\_loss).
* **Imbalance handling:** per-fold **sample weights** using inverse-frequency weighting; then weights normalized and clipped to [0.5, 3.0] to reduce extreme effects.
* **Metric tracking:** QWK computed each epoch using quadratic weights to match APTOS evaluation.

### 4.5.2. Model 2 Training Details (Stratified Split + Oversampling + Mixed Precision)

* **Split strategy:** StratifiedShuffleSplit (80/20)
  + Train = 2929, Val = 733
  + Train distribution = [1444, 296, 799, 154, 236]
  + Val distribution = [361, 74, 200, 39, 59]
* **Batch size:** 8
* **Mixed precision:** enabled (mixed\_float16) for memory efficiency at 512×512.
* **Optimizer:** AdamW (weight decay = 1e-5)
* **Warm-up stage:**
  + Backbone frozen
  + Epochs = 2
  + LR = 3e-4
* **Fine-tuning stage:**
  + Unfreeze last 40 layers only
  + Freeze BatchNormalization layers
  + Epochs = 10
  + LR = 2e-5
  + ReduceLROnPlateau + EarlyStopping (monitor val\_loss)
* **Imbalance handling:** oversampling with tf.data.Dataset.sample\_from\_datasets()
  + weights computed from inverse class counts with power scaling: OVERSAMPLE\_POWER = 0.55
  + boosts: class 1 ×1.25, class 4 ×1.15
* **Evaluation per epoch:** callback computes best cutoffs using constraints (min gap + predicted class proportion bounds) and reports the best QWK.

## **4.6Results**

### 4.6.1. Model 1 Results (5-Fold OOF QWK + Training Curve)

Because APTOS uses QWK, we report out-of-fold QWK as the main score.

**OOF QWK (naive rounding):** 0.9044021057  
**Optimized thresholds:** [0.55099074, 1.36307136, 2.58263359, 3.55118462]  
**OOF QWK (optimized thresholds):** 0.9086909720

**Classification report (OOF mapped predictions):**

* Accuracy = 0.7594
* Macro avg F1 = 0.6125
* Weighted avg F1 = 0.7749
* Best class is Grade 0 (very high precision/recall), while Grades 1 and 3 are harder due to fewer samples and similarity to neighbor grades.

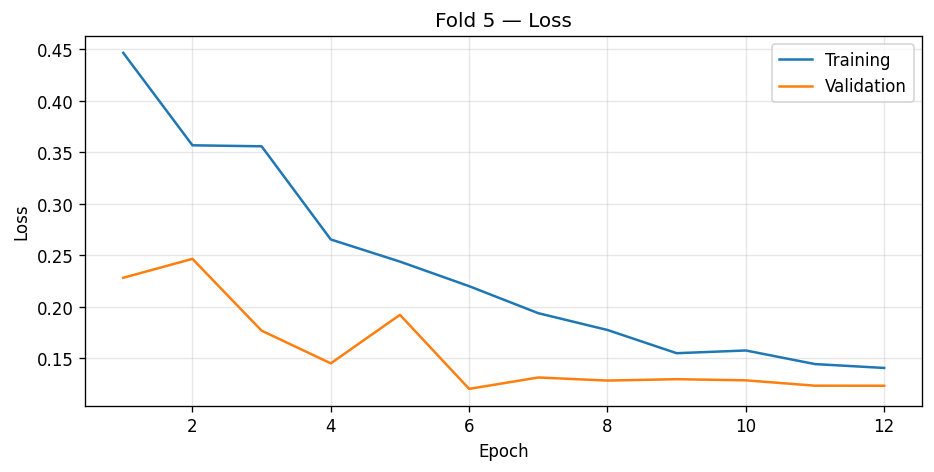


Figure 6 Model 1 Validation

### 4.6.2. Model 1 Confusion Matrix (OOF Error Analysis)

The normalized OOF confusion matrix shows:

* Strong diagonal for **grade 0** (≈0.98 correct).
* Grade 1 is often confused with grade 2.
* Grade 2 often shifts to grade 3 (and sometimes to grade 1).
* Grade 3 and grade 4 confusion exists mainly between each other (3↔4), which is expected in severity grading.

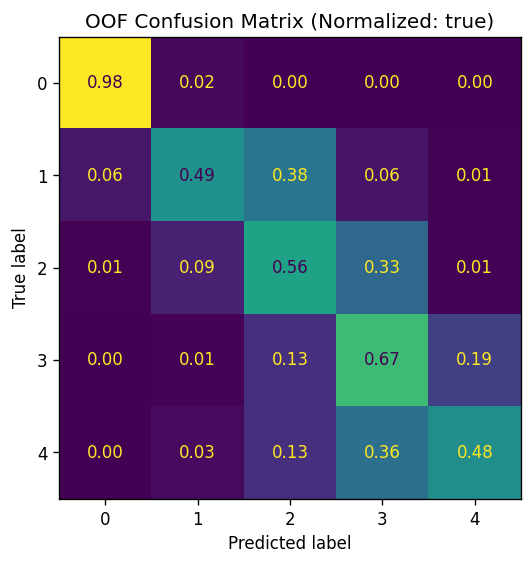


Figure 7 Model 1 Confusion Matrix

### 4.6.3. Model 2 Results (Best QWK + Training Curve)

Model 2 achieved its best validation performance using CORAL ordinal training and constrained cutoffs search.

**BEST QWK during training:** 0.9123051129  
**BEST cutoffs:** [0.7, 1.05, 2.7, 3.05]  
**FINAL QWK (best weights + best cutoffs):** 0.9123051129

**Classification report (Validation, n=733):**

* Accuracy = 0.8336
* Macro avg F1 = 0.6870
* Weighted avg F1 = 0.8321
* Grade 2 achieved strong recall (0.85), while grades 3 and 4 remain more challenging due to low support.

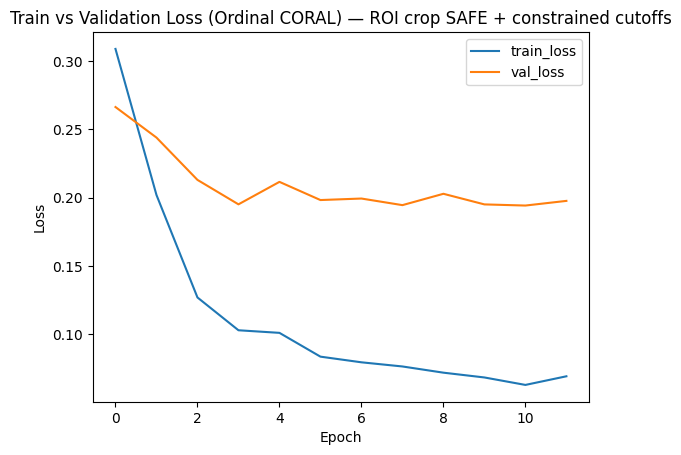


Figure 8 Model 2 Validation

### 4.6.4. Model 2 Confusion Matrix (Best QWK)

The normalized confusion matrix at best QWK shows:

* Grade 0 diagonal is strong (≈0.97 correct).
* Grade 2 diagonal is strong (≈0.85 correct).
* Grades 3 and 4 show confusion mostly with grade 2 and between each other (3↔4), which matches the ordinal difficulty.

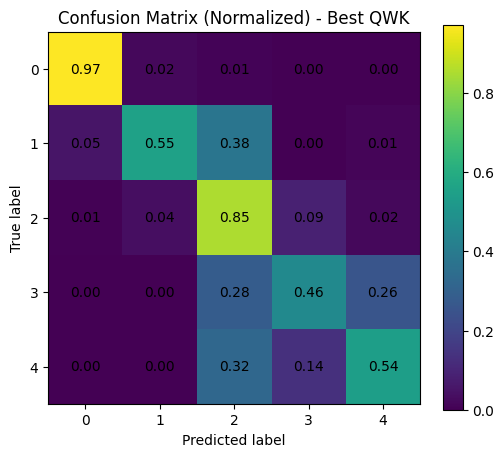


Figure 9 Model 2 Confusion Matrix

## 4.7. Comparison Between the Two Models (Results + Why Both Are Good)

### 5.7.1. Comparison Summary

Both models are designed to perform well under **QWK**, which is the official APTOS metric.

**Model 1 (Regression + CV + OptimizedRounder) is good because:**

* It uses **Stratified 5-fold CV**, which gives a reliable OOF estimate across the full dataset (better generalization estimate than a single split).
* OptimizedRounder improves QWK by learning better boundaries between grades than naive rounding.
* It provides a strong “competition-style” pipeline: OOF predictions → optimize thresholds → apply to test predictions.

**Model 2 (Ordinal CORAL + Oversampling + 512×512) is good because:**

* CORAL is a dedicated method for **ordinal labels** and encourages rank-consistent predictions.
* High resolution (512×512) + ROI crop SAFE can preserve small retinal lesions.
* Oversampling strategy directly addresses class imbalance, helping minority classes appear more during training.

### 4.7.2. Results Comparison (Important Note)

* **Model 1:** OOF QWK (optimized thresholds) = **0.90869**
* **Model 2:** Validation QWK (best cutoffs) = **0.91231**

**Important note:** these results come from different evaluation protocols:

* Model 1 uses **5-fold OOF** (covers all 3662 samples as validation across folds).
* Model 2 uses **single holdout validation split** (733 validation samples).  
  So the comparison is useful, but not a perfectly identical experimental setting.

### 4.7.3. Why Keeping Both Models Strengthens the Project

Using two different modeling formulations strengthens the project because:

* Model 1 shows that a strong transfer learning backbone + CV + threshold optimization produces stable high QWK.
* Model 2 shows that an ordinal-aware approach (CORAL) can reach strong QWK while explicitly respecting the order of grades.
* Together, they provide two valid solutions for DR grading:
  1. **Continuous severity score + optimized thresholds**
  2. **Ordinal logits + constrained cutoffs**  
     This supports the conclusion that the backbone (EfficientNetB3) is effective, and that handling the ordinal structure and imbalance can further improve results.

# Chapter 5. Research Gaps and Future Direction

## 5.1 Introduction

This chapter outlines the key research gaps identified throughout the development of our **five-grade diabetic retinopathy (DR) grading** system using the **APTOS 2019** fundus dataset. These gaps range from dataset limitations and class imbalance to model design choices, generalization under domain shift, and clinical interpretability. Addressing these gaps can improve both performance and real-world reliability, and it also opens clear future research directions for DR screening and severity grading systems.

## 5.2 Feature Representation Limited to Static Global Patterns

**Gap**  
Although CNN backbones such as EfficientNet provide strong global feature extraction, fundus-based DR grading often depends on **fine-grained local lesions** (e.g., microaneurysms and small hemorrhages). A global-only representation can miss subtle local cues, especially in early grades (mild/moderate), where discriminative signs may be small or sparsely distributed.

**Reason Not Addressed Now**  
Incorporating multi-scale lesion-aware designs (e.g., attention modules or specialized lesion branches) typically requires deeper architectural changes, higher-resolution training, and longer experimentation cycles, which were beyond the scope of the current implementation phase.

**Future Direction**  
Explore **attention-based** and **multi-scale** feature extraction (e.g., spatial/channel attention, transformer-inspired blocks, or feature pyramids) to improve sensitivity to small lesion patterns while maintaining stable performance across all five grades.

## 5.3 Class Imbalance Handling Without Broader Comparison

**Gap**  
APTOS 2019 is clearly **imbalanced across the five grades** (commonly reported as 1805/370/999/193/295 for grades 0–4). While using class/sample weights helps, it does not fully solve minority-grade underperformance, and it was not compared against other imbalance strategies (e.g., focal loss, balanced sampling, synthetic augmentation methods, or hybrid resampling).

**Reason-Not-Add-ressed-Now**  
A systematic comparison of imbalance-handling methods requires controlled experiments, consistent validation protocols, and careful tuning, particularly because different strategies can improve minority recall while harming calibration or increasing false positives.

**Future-Direction**  
Conduct an ablation study comparing **class/sample weighting** against alternative approaches (e.g., focal-style objectives, balanced batch sampling, or targeted augmentation). The goal is to improve performance on severe/proliferative grades without degrading overall QWK.

## 5.4 Metric Alignment and Ordinal Modeling Limitations

**Gap**  
APTOS 2019 evaluation is based on **Quadratic Weighted Kappa (QWK)**, which reflects ordinal distance between grades. Even when QWK is used for evaluation, training objectives often remain generic (e.g., cross-entropy or standard regression losses), which may not fully align optimization with ordinal agreement.

**Reason Not Addressed Now**  
Directly optimizing ordinal-aware objectives or differentiable kappa-style losses can introduce training instability and requires careful validation to ensure improvements generalize beyond the validation folds.

**Future Direction**  
Evaluate ordinal-specific learning formulations (e.g., ordinal regression objectives, cumulative link models, or differentiable approximations that better correlate with QWK), and compare them to regression + threshold optimization in a controlled setting.

## 5.5 Generalization Under Domain Shift and Image Quality Variability

**Gap**  
APTOS 2019 fundus images are captured under varying acquisition conditions (illumination, contrast, and artifacts). This variability can cause **domain shift**, where a model performs well on the benchmark but degrades when applied to images from different devices, clinics, or populations.

**Reason Not Addressed Now**  
Robust generalization typically requires multi-dataset training, external validation (e.g., on Messidor/EyePACS), or explicit domain adaptation—none of which were included in the current phase focused on APTOS-only implementation.

**Future Direction**  
Perform **external validation** on additional DR datasets and investigate domain-robust techniques (e.g., stain/illumination normalization, domain adaptation, test-time augmentation, or federated training where appropriate).

# Chapter 6. Conclusion

**Conclusion**

This project confirms that deep learning with transfer learning backbones can achieve strong performance for **diabetic retinopathy (DR) severity grading** on the **APTOS 2019 Blindness Detection** dataset. The implementation follows the competition setup where performance is evaluated using **Quadratic Weighted Kappa (QWK)**, which is suitable for the ordered grades (0–4).

Two complementary approaches were implemented. **Model 1 (EfficientNetB3 regression-style + threshold optimization with 5-fold CV)** achieved **OOF QWK = 0.90440** with naive rounding and improved to **0.90869** after optimized thresholds, showing that learning decision boundaries from out-of-fold predictions can improve QWK. **Model 2 (EfficientNetB3 + ordinal CORAL + constrained cutoffs, 512×512)** achieved **Best/Final QWK = 0.91231** on a stratified holdout validation split; CORAL is designed for ordinal targets and produces rank-consistent outputs.

Overall, both models are strong for different reasons: Model 1 provides robust estimation through full cross-validation and QWK-aligned thresholding, while Model 2 directly models the ordinal nature of DR grades using CORAL and benefits from high-resolution inputs and imbalance-aware sampling. Both leverage EfficientNet’s compound scaling and strong transfer learning behavior.

Key limitations are that the two scores are not perfectly comparable (CV-OOF vs single holdout), and grading remains challenging due to class imbalance and adjacent-grade confusion. Future work will evaluate both methods under the same protocol, test generalization on additional DR datasets, and explore stronger ordinal-aware objectives and attention/ensembling to reduce confusion between neighboring grades.

# Chapter 7. References

[1]. Kaggle (2019), **APTOS 2019 Blindness Detection** (competition page & dataset description). ([kaggle.com](https://www.kaggle.com/c/aptos2019-blindness-detection?utm_source=chatgpt.com))

[2]. Tan, M., & Le, Q. V. (2019), **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**, arXiv:1905.11946. ([arxiv.org](https://arxiv.org/abs/1905.11946?utm_source=chatgpt.com))

[3]. Cao, W., Mirjalili, V., & Raschka, S. (2020), **Rank consistent ordinal regression for neural networks with application to age estimation (CORAL)**, *Pattern Recognition Letters*. ([sciencedirect.com](https://www.sciencedirect.com/science/article/pii/S016786552030413X?utm_source=chatgpt.com))

[4]. Chilukoti, S. V., Shan, L., Tida, V. S., et al. (2024), **A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric**, *BMC Medical Informatics and Decision Making*. ([SpringerLink](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02446-x?utm_source=chatgpt.com))

[5]. Mohanty, C., Mahapatra, S., Acharya, B., et al. (2023), **Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy**, *Sensors*, 23(12):5726. ([mdpi.com](https://www.mdpi.com/1424-8220/23/12/5726?utm_source=chatgpt.com))

[6]. Shakibania, H., Raoufi, S., Pourafkham, B., et al. (2023), **Dual Branch Deep Learning Network for Detection and Stage Grading of Diabetic Retinopathy**, arXiv:2308.09945. ([arxiv.org](https://arxiv.org/abs/2308.09945?utm_source=chatgpt.com))

[7]. Shakibania, H., Raoufi, S., Pourafkham, B., et al. (2024), **Dual branch deep learning network for detection and stage grading of diabetic retinopathy**, *Computer Methods and Programs in Biomedicine Update / (Elsevier listing)*. ([sciencedirect.com](https://www.sciencedirect.com/science/article/abs/pii/S174680942400226X?utm_source=chatgpt.com))

[8]. Rezaee, K., & Farnami, F. (2025), **Innovative Approach for Diabetic Retinopathy Severity Classification: An AI-Powered Tool using CNN-Transformer Fusion**, (PubMed/PMC record). ([PubMed](https://pubmed.ncbi.nlm.nih.gov/40259941/?utm_source=chatgpt.com))

[9]. Rautaray, J., Ali, A. B. M., Kandpal, M., et al. (2025), **Leveraging FastViT based knowledge distillation with EfficientNet-B0 for diabetic retinopathy severity classification**, *SLAS Technology* (ScienceDirect/PubMed record). ([sciencedirect.com](https://www.sciencedirect.com/science/article/pii/S2472630325000834?utm_source=chatgpt.com))

[10]. scikit-learn Documentation, **cohen\_kappa\_score** (quadratic weighting option used for QWK evaluation). ([Scikit-learn](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html?utm_source=chatgpt.com))

[11]. TensorFlow Documentation, **tf.keras.applications.EfficientNetB3** (model backbone used in implementation). ([TensorFlow](https://www.tensorflow.org/api_docs/python/tf/keras/applications/EfficientNetB3?utm_source=chatgpt.com))

[12]. TensorFlow Documentation, **tf.keras.optimizers.AdamW** (optimizer used in Model 2). ([TensorFlow](https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/AdamW?utm_source=chatgpt.com))

[13]. scikit-learn Documentation, **StratifiedKFold** (used for stratified 5-fold CV in Model 1). ([Scikit-learn](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html?utm_source=chatgpt.com))