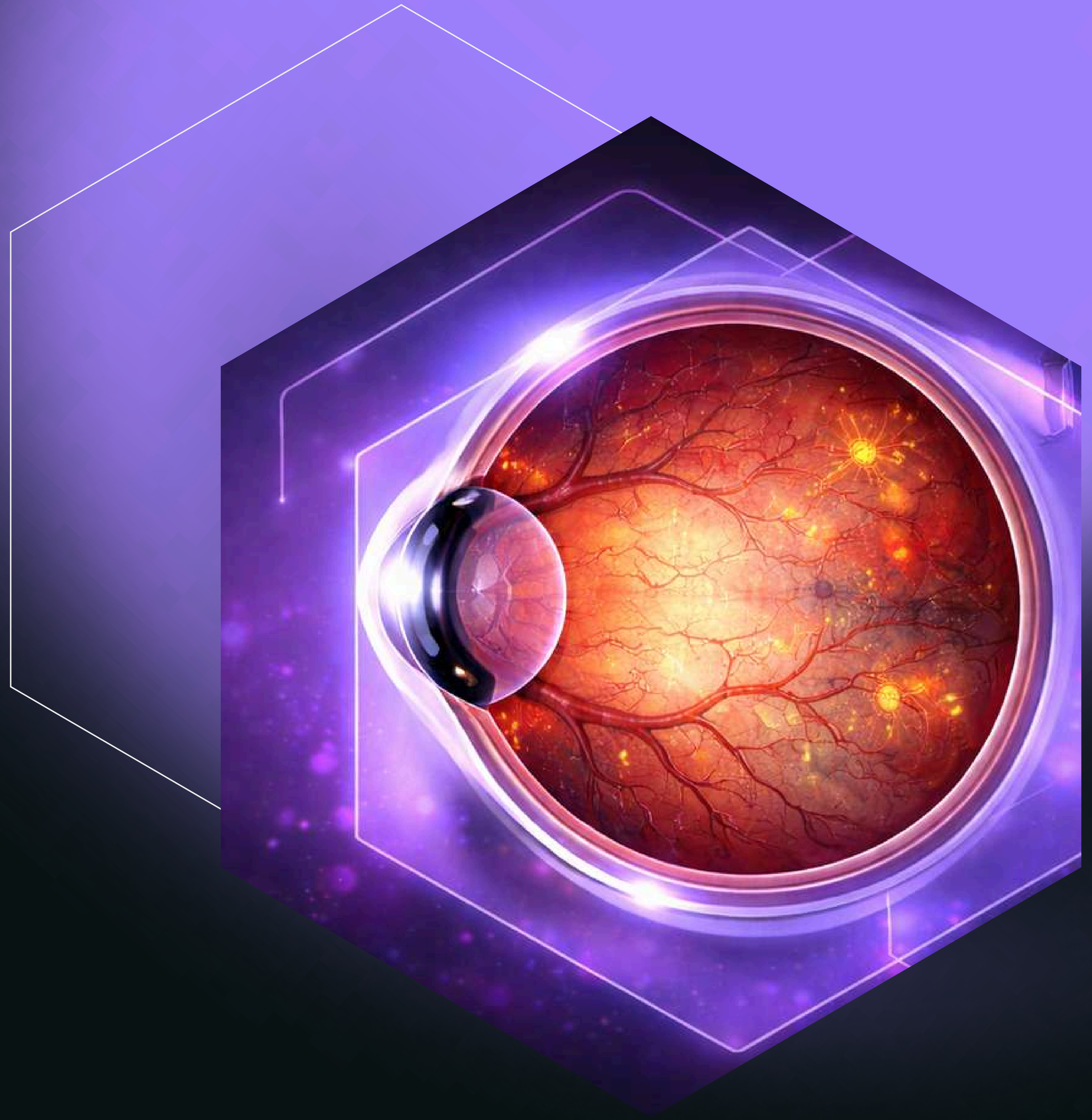
Improving Robustness and Interpretability of Deep Learning Models for Multi-Stage Diabetic Retinopathy Classification under Low-Resource Imaging Conditions



Gitanjali Atri, , Lavanya Gupta, Sukant

20th May, 2026

The Problem

A silent but growing complication of diabetes

Leading microvascular complication of diabetes mellitus.
Major cause of preventable blindness in working-age adults.

Global Burden

- 103 million people affected in 2020
- Projected 130 million by 2030
- Expected 161 million by 2045
- >25% increase within 10 years

Why the Rise?

- Increasing global diabetes prevalence
- Lifestyle changes
- Longer life expectancy & aging populations

India Burden

3 million

people aged ≥40 years have VTDR

Individuals with known diabetes

Higher Risk Groups

Higher prevalence in high- and middle-ETI-SDI states

Geographic Pattern

Tan, T.-E., & Wong, T. Y. (2023). Diabetic retinopathy: Looking forward to 2030. Frontiers in Endocrinology, 13, 1077669. https://doi.org/10.3389/fendo.2022.1077669

Landscape of ML for DR Screening – Global & India

Global Progress in AI for DR

- **Ophthalmology has been a pioneer in clinical AI development**, with automated DR detection from color fundus photographs (CFP) being one of the earliest clinical AI applications (since 2016).
- Early large-dataset studies demonstrated high diagnostic accuracy for detecting referable DR and vision-threatening DR (VTDR).

Nearly 70%

of India's population lives in rural areas, where access to ophthalmologists is limited.

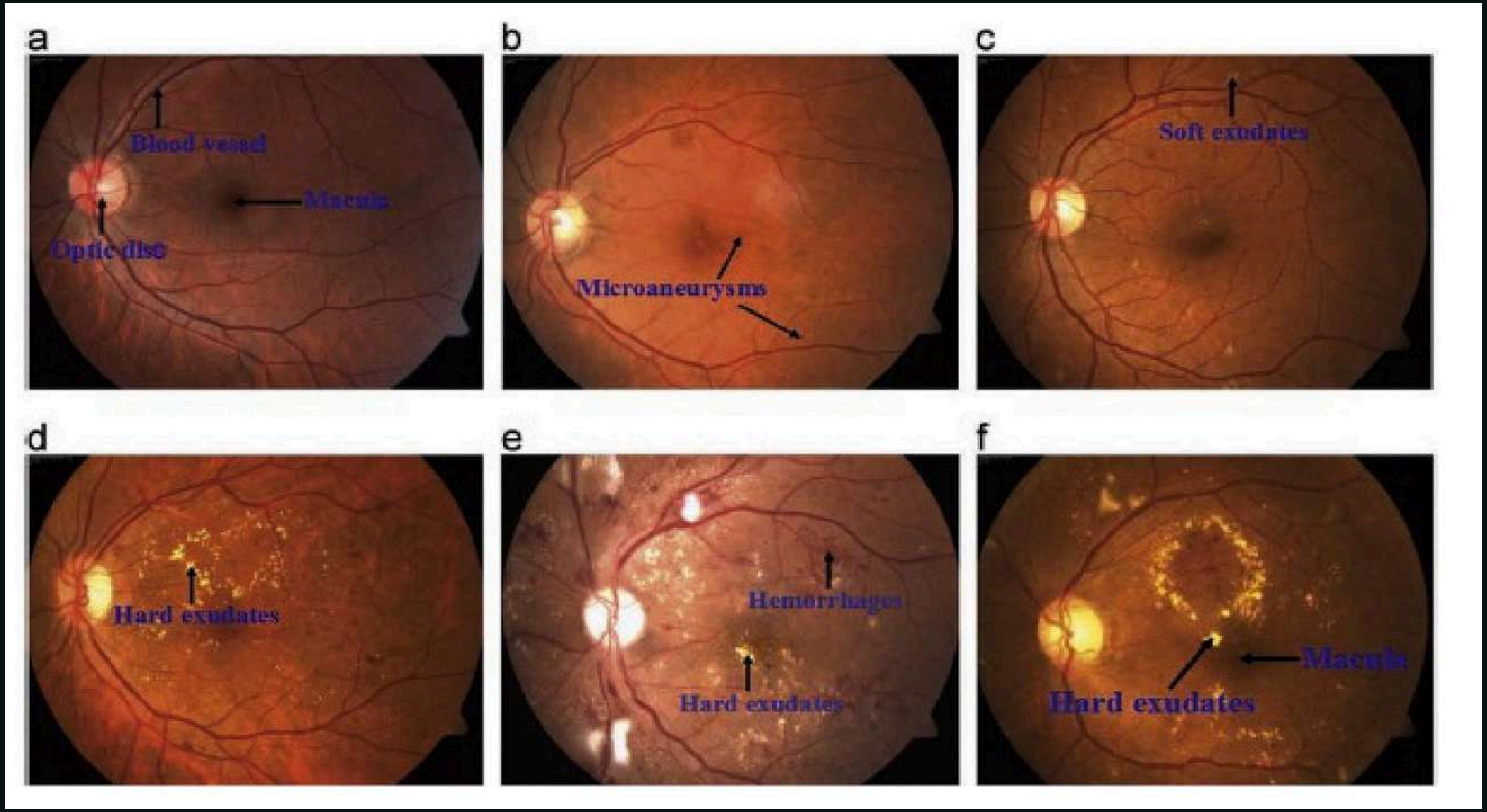
Challenges in Rural Screening Settings

- Rural screening programs face practical challenges, including patient movement during imaging, long capture times (10–12 minutes per patient), and limited device battery in field conditions.
- AI models trained on high-quality clinical images may underperform in rural screening environments, where image quality is often variable due to field conditions and limited imaging infrastructure.

AI trained on fundus images achieved a sensitivity of 90.3% and a specificity of 98.1% in detecting referable diabetic retinopathy without the need for the advanced technology of OCT.

AI can recognize important indicators of diabetic retinopathy such as, microaneurysms, hemorrhages, and exudates based solely on fundus image making it highly efficient in broad clinical use

Development and Validation of a Deep Learning Algorithm forDetection of Diabetic Retinopathy in Retinal Fundus Photographs



Macula: The central region of the retina responsible for sharp, detailed central vision.

Microaneurysms (MAs): Small red dots representing localized capillary dilations and the earliest clinically visible sign of diabetic retinopathy.

Hemorrhages (HMs): Retinal bleeding appearing as blot or flame-shaped red lesions, often caused by rupture of microaneurysms or damaged vessels.

Hard Exudates (HEs): Yellow-white lipid deposits in the retina resulting from chronic vascular leakage, frequently located near the macula.

Soft Exudates (Cotton Wool Spots): Pale, fluffy white lesions caused by localized retinal nerve fiber layer ischemia.

Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. Informatics in Medicine Unlocked, 20, 100377. https://doi.org/10.1016/j.imu.2020.100377

Impact of Our Solution

Improved Reliability in Real-World Settings

Enhances the robustness of deep learning models to accurately analyze low-quality or artifact-affected fundus images commonly encountered in low-resource clinical environments.

Better Clinical Trust through Interpretability

Provides interpretable model outputs, allowing clinicians to understand which retinal features contribute to AI predictions, increasing confidence in automated diagnosis.

Multi-Stage DR Classification

Enables reliable detection and classification across multiple stages of diabetic retinopathy, supporting both early diagnosis and identification of advanced disease.

Scalable Screening in Low-Resource Settings

Supports large-scale automated screening, helping extend diabetic retinopathy detection to resource-limited healthcare systems.

Literature Review

Deep learning (DL) models have become the primary approach for automated diabetic retinopathy (DR) screening using retinal fundus images.

- The most comprehensive systematic synthesis of DL research for DR screening was conducted by Wu et al. and published in the Journal of Medical Internet Research. The study performed a **meta-analysis of 60 research articles** collected from major databases, including PubMed and EMBASE, covering literature published from database inception until June 2020, providing one of the most rigorous evaluations of DL-based diabetic retinopathy screening systems.
- A more recent study by Moannaei et al. analyzed **76 studies covering 1.37 million retinal images**, reporting a pooled sensitivity of 90.54% and specificity of 78.33%, demonstrating high diagnostic capability.
- The lower pooled specificity compared to earlier reviews reflects the increasing use of real-world datasets and lower-quality retinal images, highlighting that model performance often degrades outside controlled benchmark conditions.

Wu, J.-H., et al. (2021). Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis. Journal of Medical Internet Research, 23(7), e23863. https://doi.org/10.2196/23863

Moannaei, et al. (2025). Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: A meta-analysis. BioMedical Engineering OnLine, 24, 34. https://doi.org/10.1186/s12938-025-01336-1

Current Approaches DR Detection

- Deep learning models based on convolutional neural networks (CNNs) have demonstrated strong performance in diabetic retinopathy detection.
- Architectures such as ResNet, EfficientNet, and Vision Transformer achieve high classification accuracy across public datasets including EyePACS, APTOS, and MESSIDOR, often outperforming traditional image processing techniques in terms of sensitivity and AUC.

Several AI and DL methodologies utilized are:

- **Deep CNNs** for automated retinal feature extraction and DR classification.
- **Vision Transformers (ViTs)** that use self-attention mechanisms to capture global spatial relationships in high-resolution retinal images.
- **Recurrent Neural Networks (RNNs)** for temporal feature analysis and sequential lesion tracking.
- Some studies also propose the use of optimization algorithms, such as **Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC)**, to improve model performance by optimizing model parameters and enhancing convergence during training.
- Hybrid AI models have been increasingly proposed, **combining CNNs with classical machine learning classifiers (e.g., support vector machines, decision trees)**, These hybrid approaches not only enhance lesion localization and classification accuracy but also improve generalization across diverse clinical datasets.

Alshammari, R. D., Abu Haliqa, W. M., & Majzoub, R. Y. (2025). Diagnostic accuracy of AI- or DL-enhanced technologies in the diagnosis of diabetic retinopathy: A systematic review. Journal of Medical Imaging and Interventional Radiology, 12, 39. https://doi.org/10.1007/s44326-025-00083-w

Research Gaps

Where current DR detection models fall short

Based on our Literature Review

Lack of Robust External Validation

Moannaei, M., Jadidian, F.,et al. (2025). *Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: A meta-analysis.* Discover Applied Sciences

Limitations of Binary DR Classification

Bhulakshmi D, Rajput DS. 2024. *A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. PeerJ Computer Science 10:e1947*

AI Model Performance Degradation in Low-Resource Settings

Duggal, M., et al. (2025). Real-world evaluation of AI-driven diabetic retinopathy screening in public health settings: Validation and implementation study. JMIR Medical Informatics, 13, e67529.

Limited Indian Data in Existing DR Models

Raman, R., et al. (2022). *Prevalence of diabetic retinopathy in India stratified by known and undiagnosed diabetes, urban–rural locations, and socioeconomic indices: Results from the SMART India population-based cross-sectional screening study.* The Lancet Global Health

Lack of Quantitative Validation for Model Interpretability

Lepetit-Aimon, G.,et al. (2024). *MAPLES-DR: MESSIDOR anatomical and pathological labels for explainable screening of diabetic retinopathy.* Scientific Data, 11, 914.

Loss Function: CE or OD?

Afshan Hashm (2026). *Robust diabetic retinopathy grading using dual-resolution attention-based deep learning with ordinal regression. arXiv.*

Mensah, S. O., Bah, B., & Brink, W. (2021). Towards the localisation of lesions in diabetic retinopathy. arXiv.

Datasets Description

Overview of the Datasets Used

APTOS 2019 Blindness Detection

APTOS 2019 Blindness Detection

Nature of Dataset	• Retinal fundus images labelled by diabetic retinopathy severity. • ~3,662 images with five severity levels (0–4).
Reason for Selection	• Real-world, well-labeled dataset suitable for machine learning training and evaluation. • Graded by trained ophthalmologists for clinical reliability.
Data Collection	• Collected by Aravind Eye Hospital, Tamil Nadu for APTOS 2019 Blindness Detection Competition. • Images captured in hospitals and eye clinics using fundus cameras.
Ethical concerns	• Most images from rural India, where eye screening access is limited. • Potential geographic bias; may require further validation for other populations.

EyePACS Diabetic Retinopathy Detection Dataset

EyePACS Diabetic Retinopathy Detection Dataset

Nature of Dataset	• ~ Contains 35,126 color-fundus images, each is of size 3888 ×2951. • Images labelled based on disease severity levels (0–4).
Reason for Selection	•Large, real-world dataset appropriate for machine learning research and model development. • Labels verified by trained clinicians.
Data Collection	• Images were collected through the EyePACS DR screening program. • Captured using digital fundus cameras in real-world clinical environments.
Ethical concerns	• Potential dataset bias due to limited demographic and geographic representation. • Variability in image quality caused by differences in lighting conditions, camera types, and resolution. • Models trained on this dataset may require additional validation.

Indian Diabetic Retinopathy Image Dataset (IDRiD).

Indian Diabetic Retinopathy Image Dataset (IDRiD)

Nature of Dataset	• Includes both disease grading and detailed lesion-level annotations (e.g., microaneurysms, hemorrhages, exudates). • Comprises 516 retinal fundus images.
Reason for Selection	• Used as an independent dataset for external validation. • Enables evaluation of model generalization to unseen data. • Provides high-quality expert annotations for both grading and lesion segmentation tasks.
Data Collection	• Developed by the Indian Institute of Technology (IIT) Madras in collaboration with Sankara Nethralaya. • Images captured using fundus cameras in real-world clinical settings. • Data collected from diabetic patients in India. • Ophthalmologists graded images and annotated lesions following clinical standards.
Ethical concerns	• Potential population bias, as images are primarily from Indian patients. • Relatively small dataset size, which may limit model generalizability. • Clinical validation required before applying models to broader and more diverse populations.

MESSIDOR-2 Dataset

Messidor-2

Nature of Dataset	• Images are labeled based on the presence and severity of diabetic retinopathy. • Comprises 1,748 retinal fundus images.
Reason for Selection	• Employed as an independent dataset for external validation. • • Assesses model performance on unseen clinical data. • • Useful for evaluating model robustness and cross-dataset generalization.
Data Collection	• Dataset provided by ADCIS, France. • Retinal images captured using fundus cameras in clinical settings. • Images graded and reviewed by ophthalmologists.
Ethical concerns	• Potential population bias due to data originating from specific clinical centers. • Models trained on different datasets may perform variably on this population. • Careful validation is required before applying models to diverse global populations

Dataset Name	Grade 0 (no DR)	Grade 1 (mild)	Grade 2 (moderate)	Grade 3 (severe)	Grade 4 (proliferative)	Total
Aptos	1805	370	999	193	295	3662
IDRiD	504	75	504	279	186	1548
EyePACS	25,810	2443	5292	873	708	35,126
Messidor -2	1017	270	156	191	75	1709

Data Acquisition and Features Preprocessing

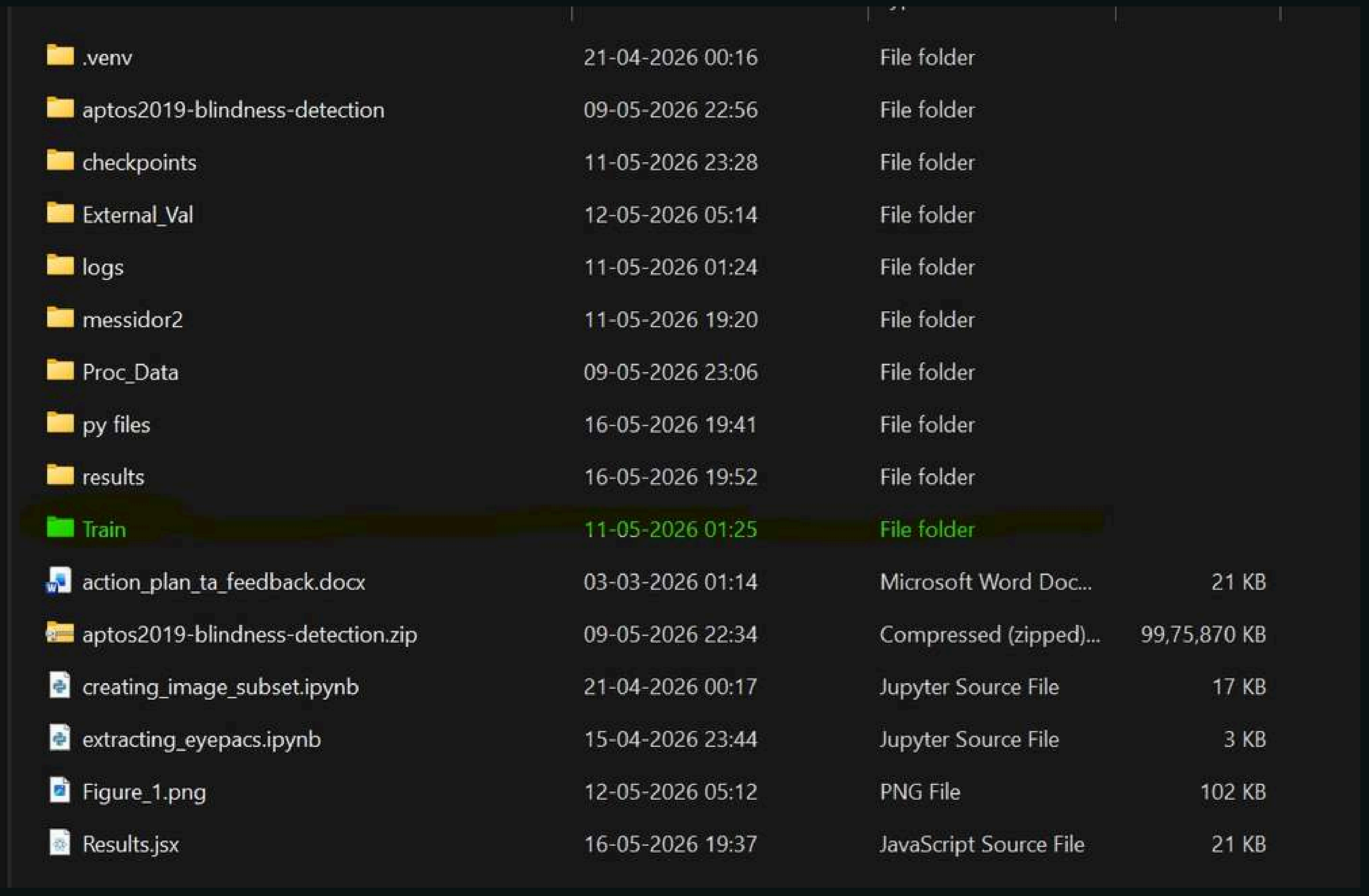
Dataset Collection

Training Datasets

- EyePACS (15,162 images)
- APTOS 2019 Blindness Detection (3,662)

These datasets were combined to improve:

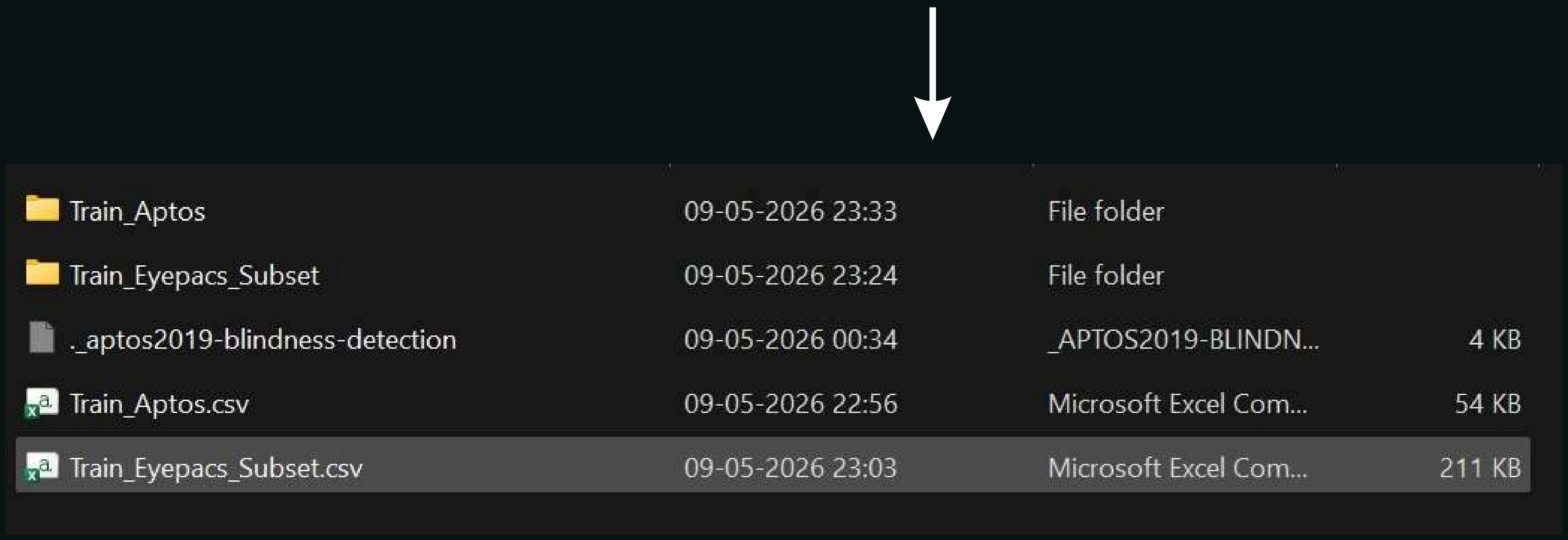
- data diversity,
- class distribution,
- and model generalization across acquisition conditions.

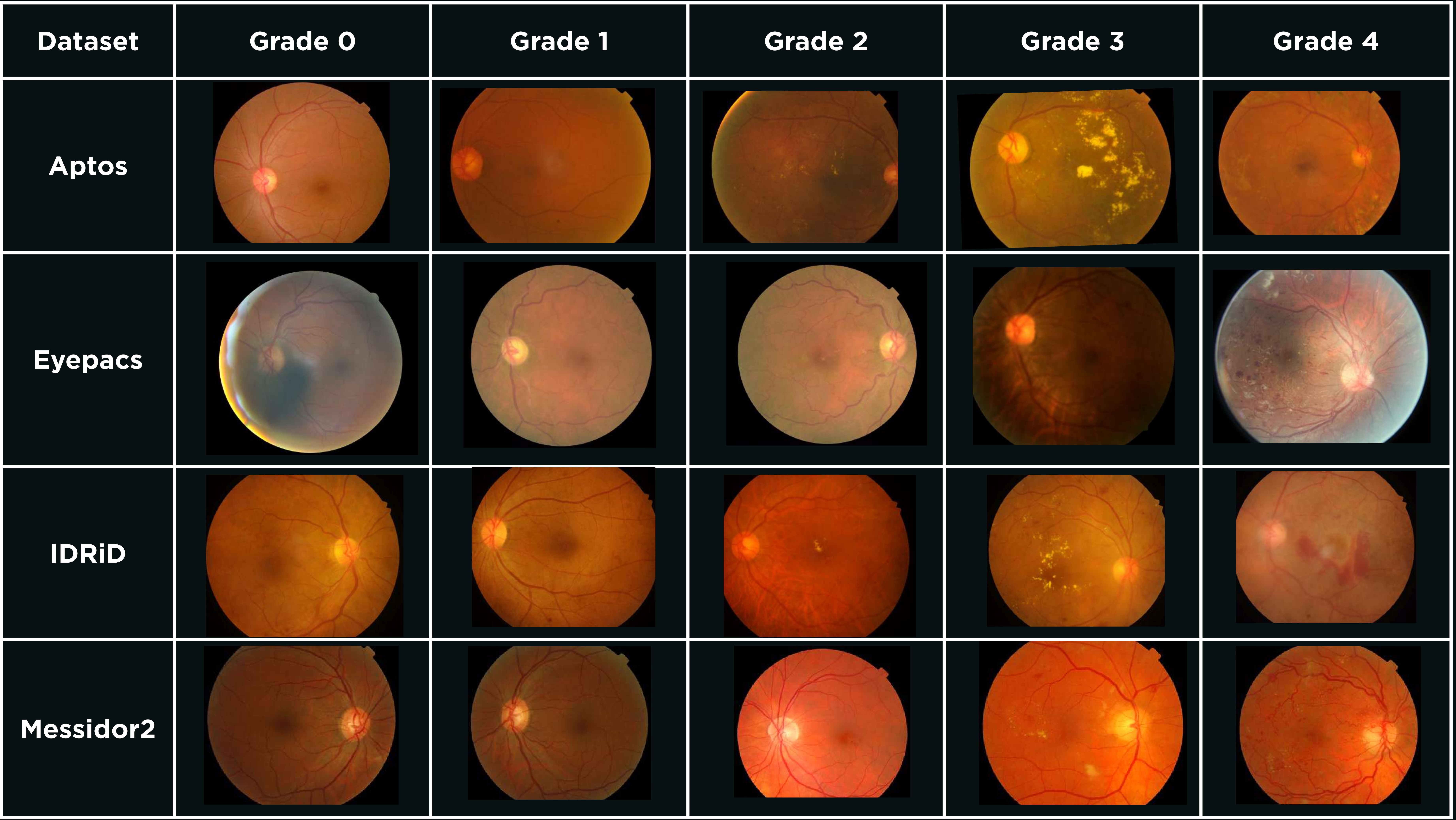


External Validation Datasets

- IDRiD
- Messidor-2

These datasets were reserved strictly for external evaluation and never used during training.




Dataset
Grade 0
Grade 1
Grade 2
Grade 3
Grade 4
Aptos
Eyepacs
IDRiD
Messidor2

Dataset Splitting

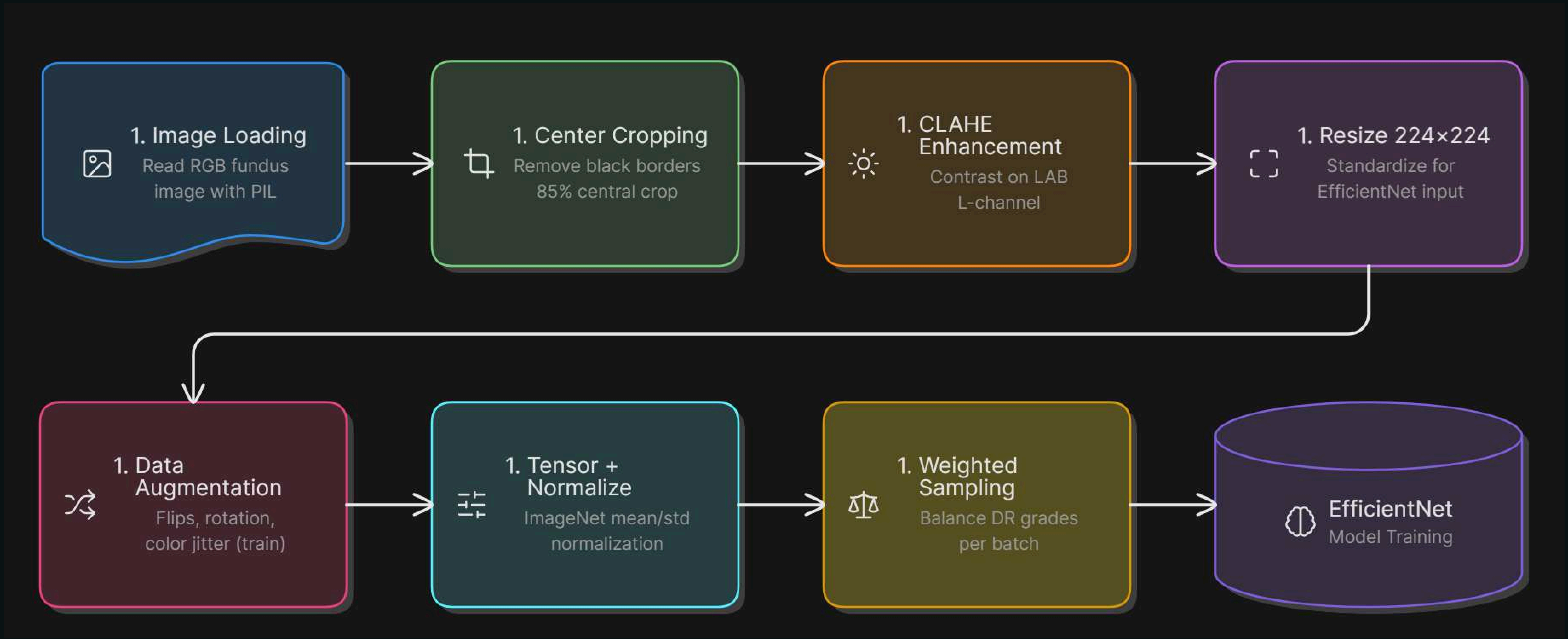
The combined EyePACS + APTOS dataset was split (80-20) into:

- Training set,
- validation set.

The split preserved class distribution to reduce imbalance effects.

External datasets were never included in optimization or hyperparameter tuning.

Fundus Images Preprocessing Pipeline



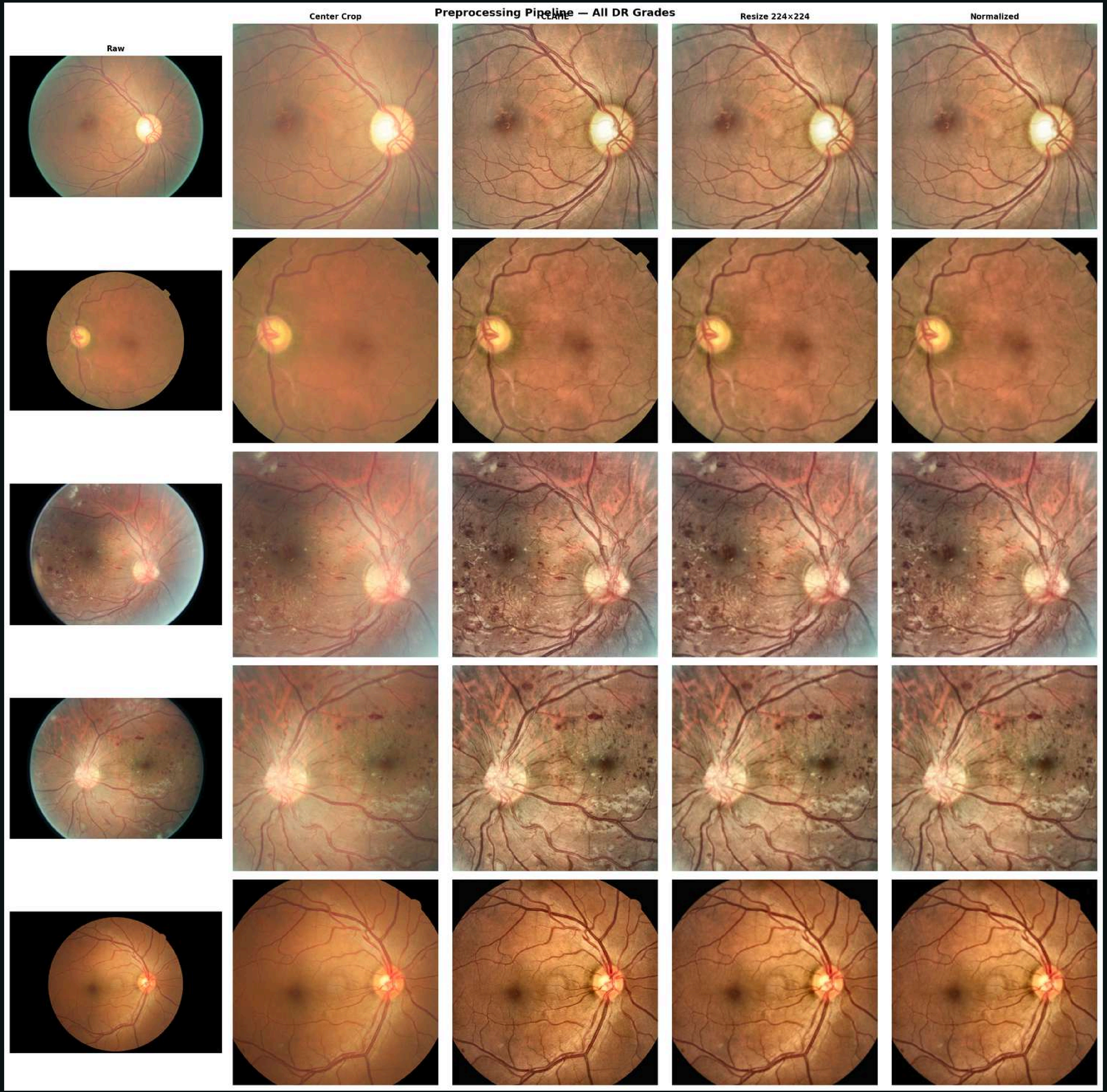
0

1

2

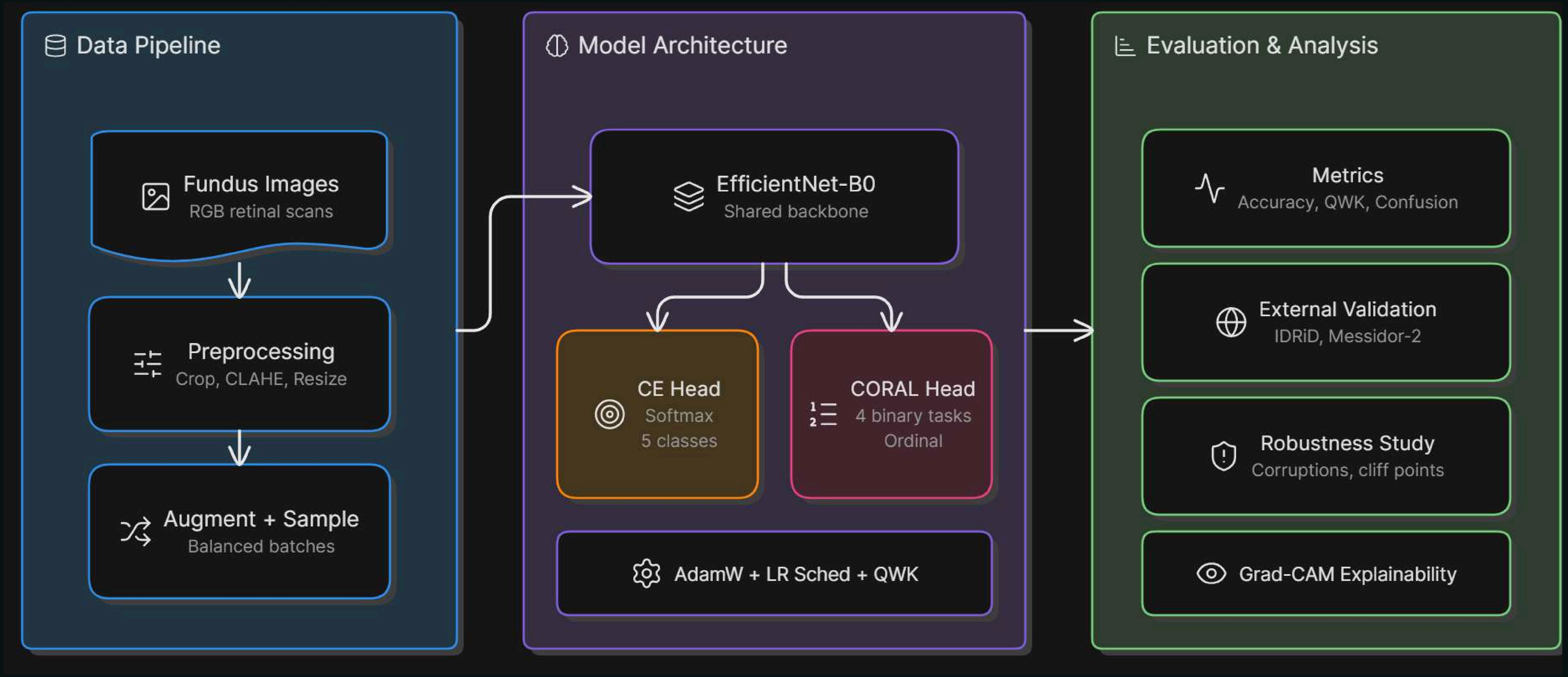
3

4


Preprocessing Pipeline — All DR Grades
Raw
Center Crop
CLAHE
Resize 224×224
Normalized

Methodology

Methodology



Baseline Model Architecture

EfficientNet-B0 pretrained on ImageNet.

The final classification layer was replaced with **nn.Linear(in_features, 5)** to predict the five DR severity grades.

Dharrao, D., Dharrao, M., Patil, S., Salvin, S., Ahire, P., & Dongre, Y. (2025). AI-driven detection and classification of diabetic retinopathy stages using EfficientNetB0. Discover Applied Sciences, 7, 1400. https://doi.org/10.1007/s42452-025-07998-9

Loss Function Experiments

Two training paradigms were implemented.

- **Cross Entropy Model (CE)** where all classification errors are treated equally.
- **Ordinal Loss Model (OD)** which penalizes distant-grade mistakes more strongly.

Optimization Strategy

Training used: AdamW

AdamW decouples weight decay from gradient updates, improving:

- regularization,
- convergence stability,
- and generalization.

Initial Training Configuration

Initial experiments used:

- BATCH_SIZE = 8
- LEARNING_RATE = 1e-4
- LR_STEP_SIZE = 2
- LR_GAMMA = 0.5

Training instability was observed:

- noisy gradients,
- fluctuating validation metrics,
- premature LR decay.

Optimization Refinement

Final Configuration

- BATCH_SIZE = 32
- LEARNING_RATE = 3e-4
- LR_STEP_SIZE = 4
- LR_GAMMA = 0.3

- Improved gradient stability and reduced stochastic variance.
- Scaled proportionally with batch size to preserve optimization dynamics.
- Prevented premature learning slowdown during active feature learning stages.
- Allowed finer optimization during late-stage convergence.

Learning Rate Scheduling

Training used: StepLR

The scheduler progressively reduced LR during training to:

- stabilize convergence,
- refine parameter updates,
- improve late-stage performance.

Model Selection Criterion

Models were not selected using validation loss.
Instead, checkpoints were saved using:
Quadratic Weighted Kappa (QWK)

- QWK is the standard metric for DR grading,
- It captures ordinal agreement,
- Better reflects clinical grading quality.

Best-performing checkpoints were automatically stored.

Evaluation Metrics

Primary Metric

- Quadratic Weighted Kappa (QWK)

Secondary Metrics

- Validation accuracy
- Validation loss
- Confusion matrix
- Per-class classification metrics

Quadratic Weighted Kappa (QWK) is a statistical metric that measures the agreement between two raters or systems (such as human graders and AI) on an ordinal scale, where the distance between categories matters.

Internal Validation

Both CE and ordinal models were evaluated on **held-out validation data(80-20).**

Analysis included:

- Epoch-wise QWK progression,
- Training stability,
- Convergence behavior,
- Class confusion analysis.

External Validation

Generalization performance was evaluated on:

- **IDRiD**
- **Messidor-2.**

Separate external evaluation scripts loaded:

- **Trained CE model**
- **Trained ordinal model**

and computed:

- QWK,
- accuracy,
- confusion matrices,
- classification reports.

Robustness Evaluation

A degradation study was performed to evaluate performance under image corruption.
Synthetic degradations included:

- **Blur**
- **Noise**
- **Brightness shifts**
- **Contrast degradation**

Each degradation was applied progressively at increasing severity levels.

For every severity level:

- QWK was measured
- Degradation curves were plotted
- Robustness cliffs were identified.

Comparison focused on:

- OD model robustness
- CE model robustness

Explainability Analysis

Model interpretability was analyzed using: Grad-CAM

GradCAM heatmaps were generated for:

- Both CE and ordinal models, across all DR grades.

Purpose:

- Visualize retinal regions influencing predictions
- Compare lesion localization behavior
- Evaluate whether ordinal training improves clinical attention alignment

Comparative Analysis

Final analysis **compared CE and ordinal models** across:

Aspect	Comparison
Internal QWK	CE vs OD
External QWK	CE vs OD
Robustness	CE vs OD
Degradation tolerance	CE vs OD
Attention localization	CE vs OD

Literature Review

Mensah, S. O., Bah, B., & Brink, W. (2021). Towards the localisation of lesions in diabetic retinopathy. arXiv. https://arxiv.org/abs/2012.11432

- Trained and evaluated using the APTOS 2019 Blindness Detection dataset
- Transfer learning was implemented using pre-trained CNN architectures—VGG16, ResNet50, InceptionV3, and InceptionResNetV2

Loss Function: Categorical Cross-Entropy Loss

Limitations

- The study mainly focused on visualization and explainability rather than detailed clinical validation.

Performance

- Best-performing model: InceptionV3
- Achieved accuracy ≈ 96.07%

Afshan Hashm (2026). Robust diabetic retinopathy grading using dual-resolution attention-based deep learning with ordinal regression arXiv. https://arxiv.org/abs/2604.17341

- Trained and evaluated using large-scale retinal fundus image datasets (e.g., APTOS / EyePACS DR datasets)
- Integrated an attention mechanism to focus on clinically relevant regions

Used Ordinal Regression formulation instead of standard classification to respect DR severity ordering

Limitations

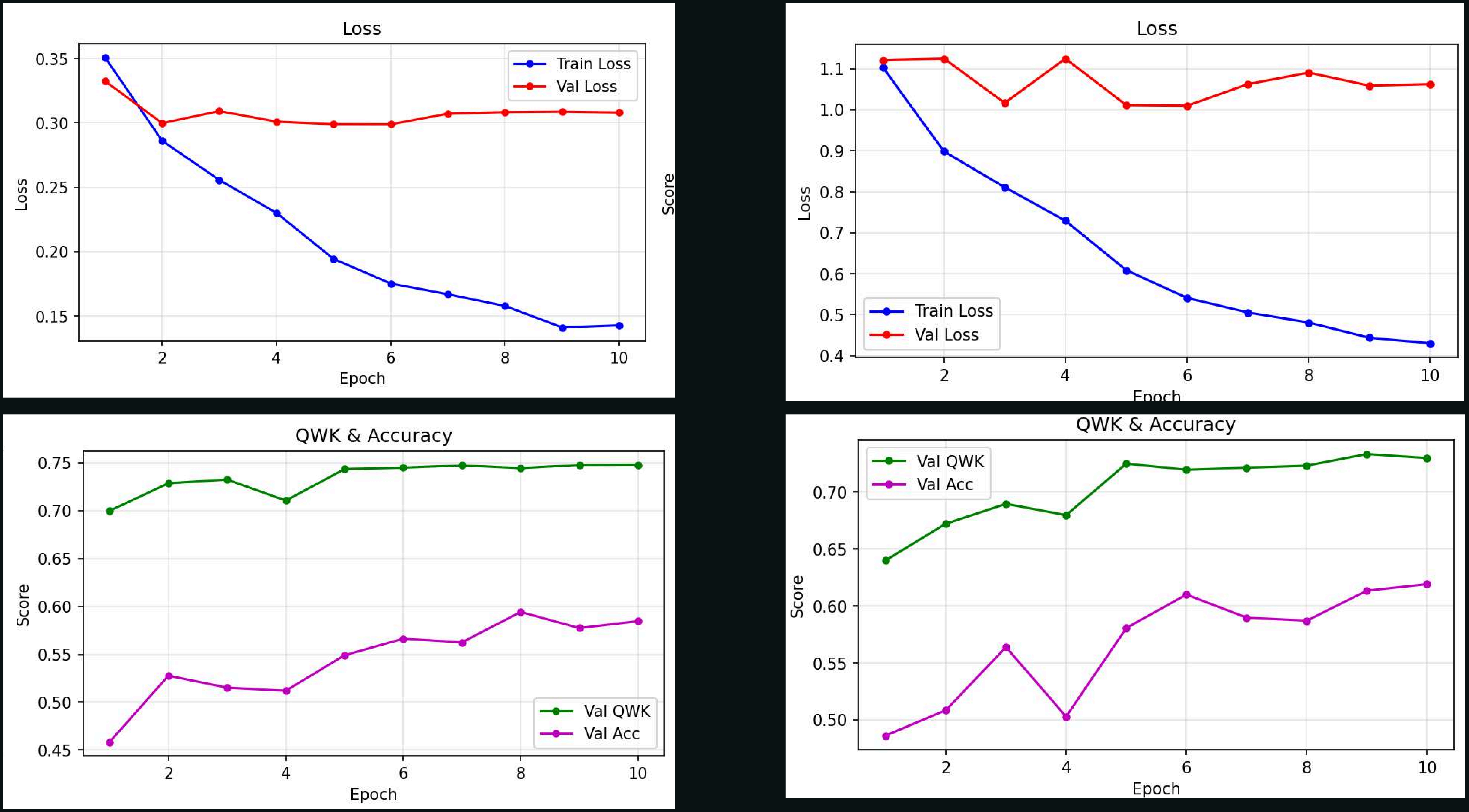
- Requires more training time and GPU memory
- Performance depends on proper ordinal threshold learning

Performance

- Quadratic Weighted Kappa (QWK): ~0.88 – 0.91
- Accuracy: ~82% – 86% (depending on dataset split)

Results

ORD Training Progression CE



Ordinal Framework (ORD)

Achieves optimal validation loss at ~0.30 with a smooth, non-volatile curve. Training and validation tracks closely, indicating superior generalization without overfitting.

Peak QWK: 0.75 (Epoch 10)

Cross-Entropy (CE)

Exhibits high volatility in validation loss, ending at ~1.05. The disparity between training and validation suggests a fractured decision boundary for ordered labels.

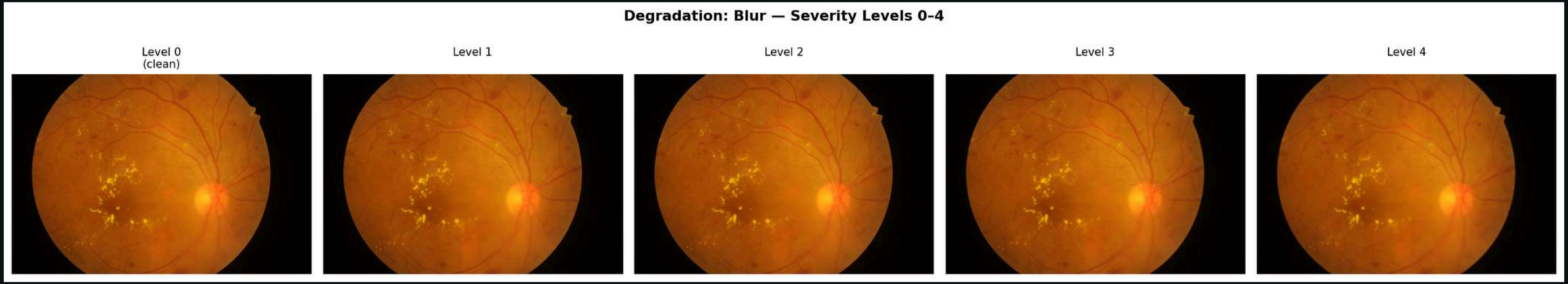
Peak QWK: 0.73 (High Volatility)

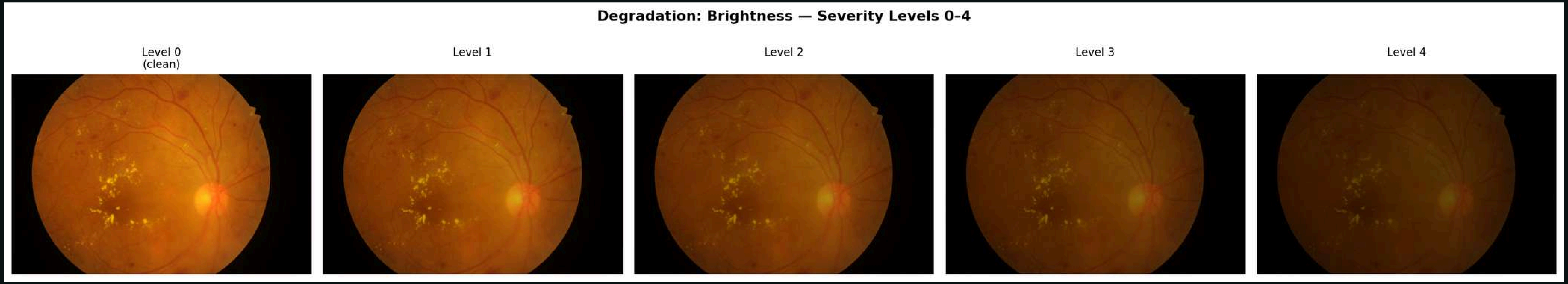
Cross-Dataset Generalization

model	dataset	accuracy	qwk
ordinal	idrid	0.538461538	0.825845375
ordinal	messidor2	0.669151376	0.759352113
ce	idrid	0.56043956	0.806135492
ce	messidor2	0.701261468	0.738666448

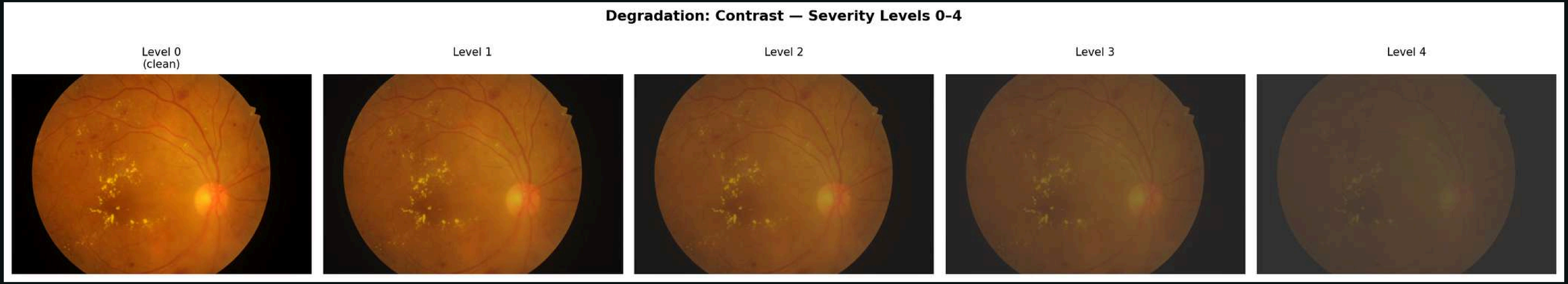
The ordinal model consistently yields superior QWK scores over CE on both **IDRiD (0.83 vs. 0.81)** and **Messidor2 (0.76 vs. 0.74)**, despite CE achieving higher raw accuracy on both datasets **(0.56 and 0.70 respectively).**

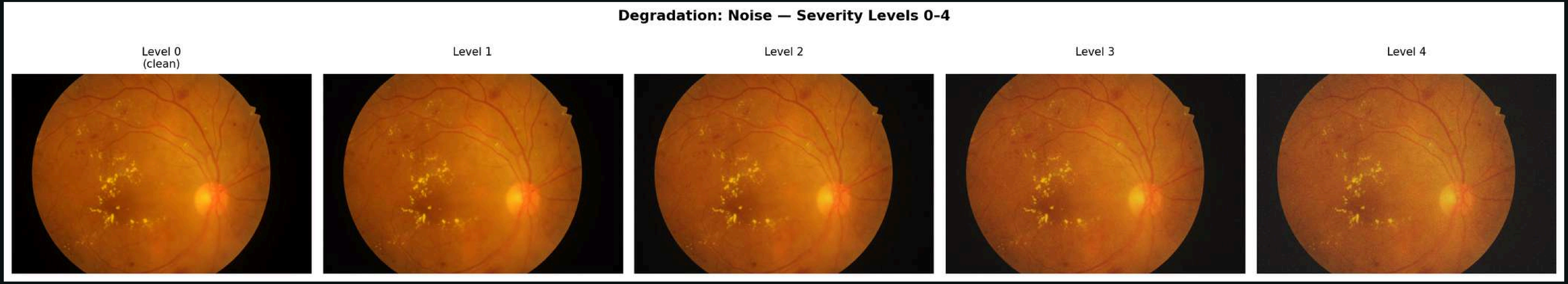
Robustness Under Degradation



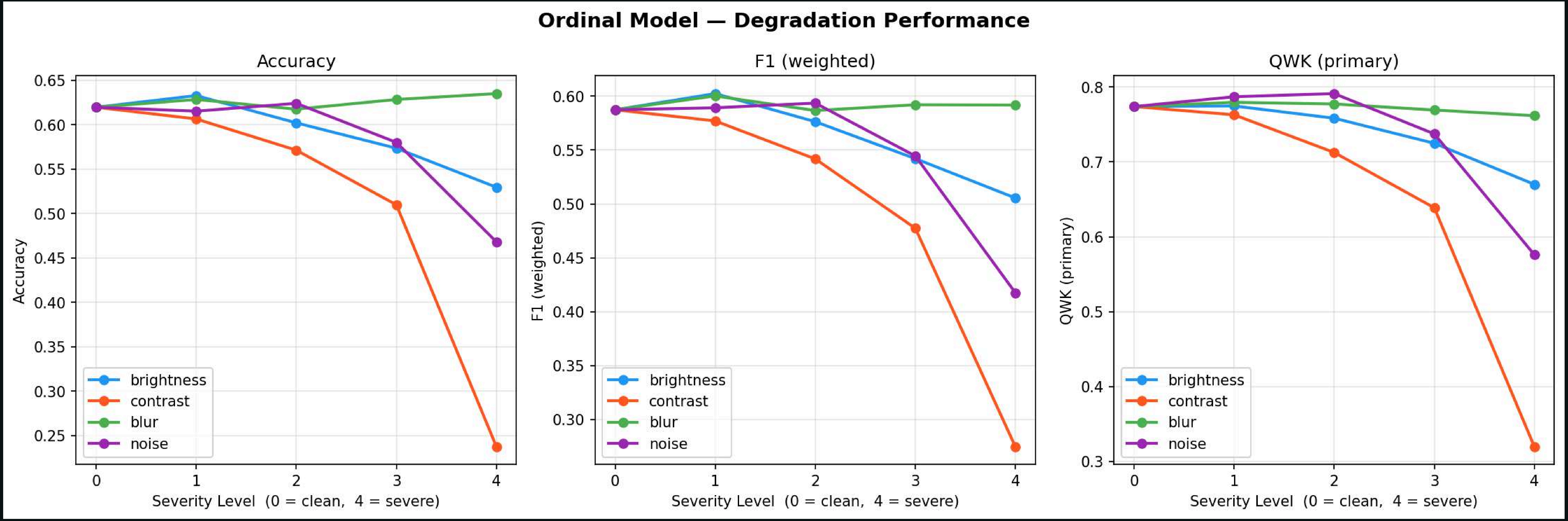


Robustness Under Degradation

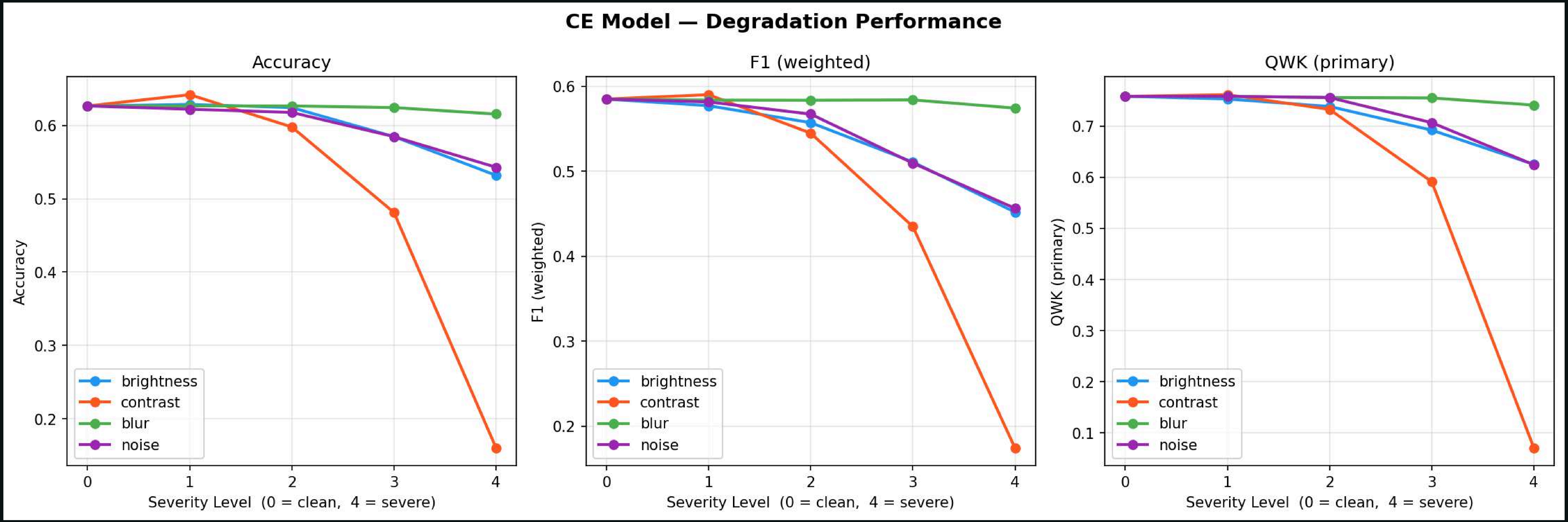




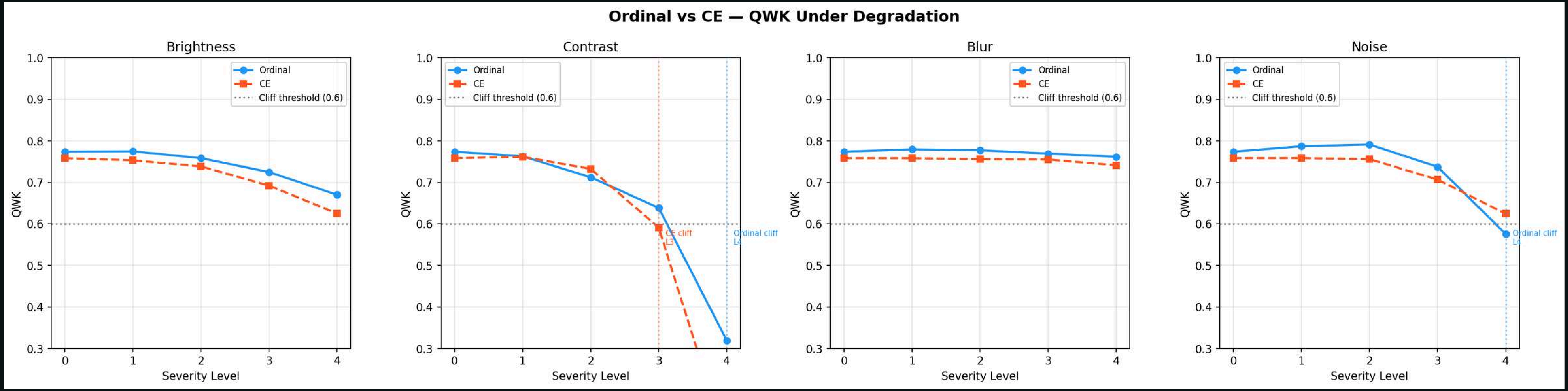
Robustness Under Degradation



Robustness Under Degradation



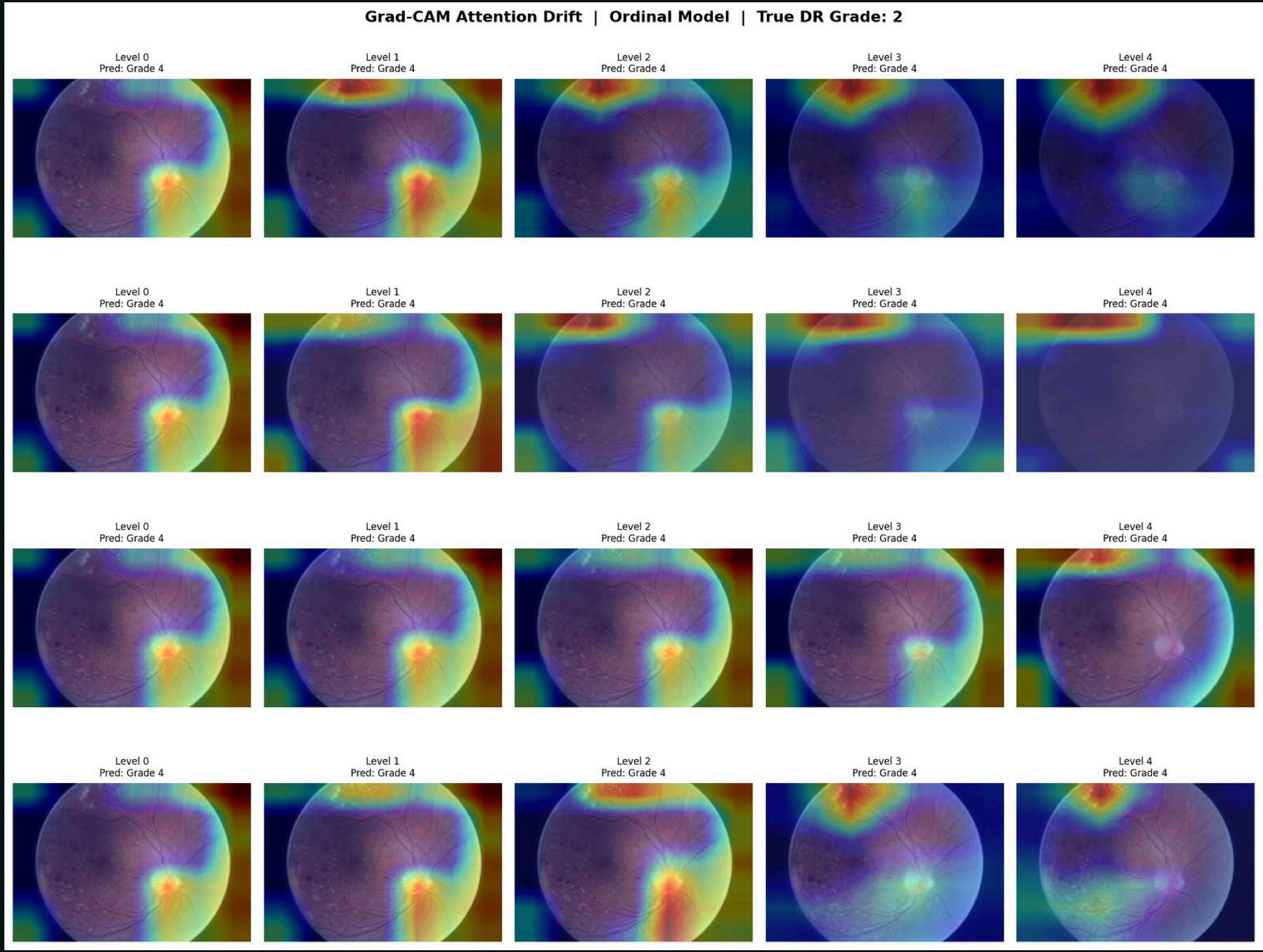
Robustness Under Degradation



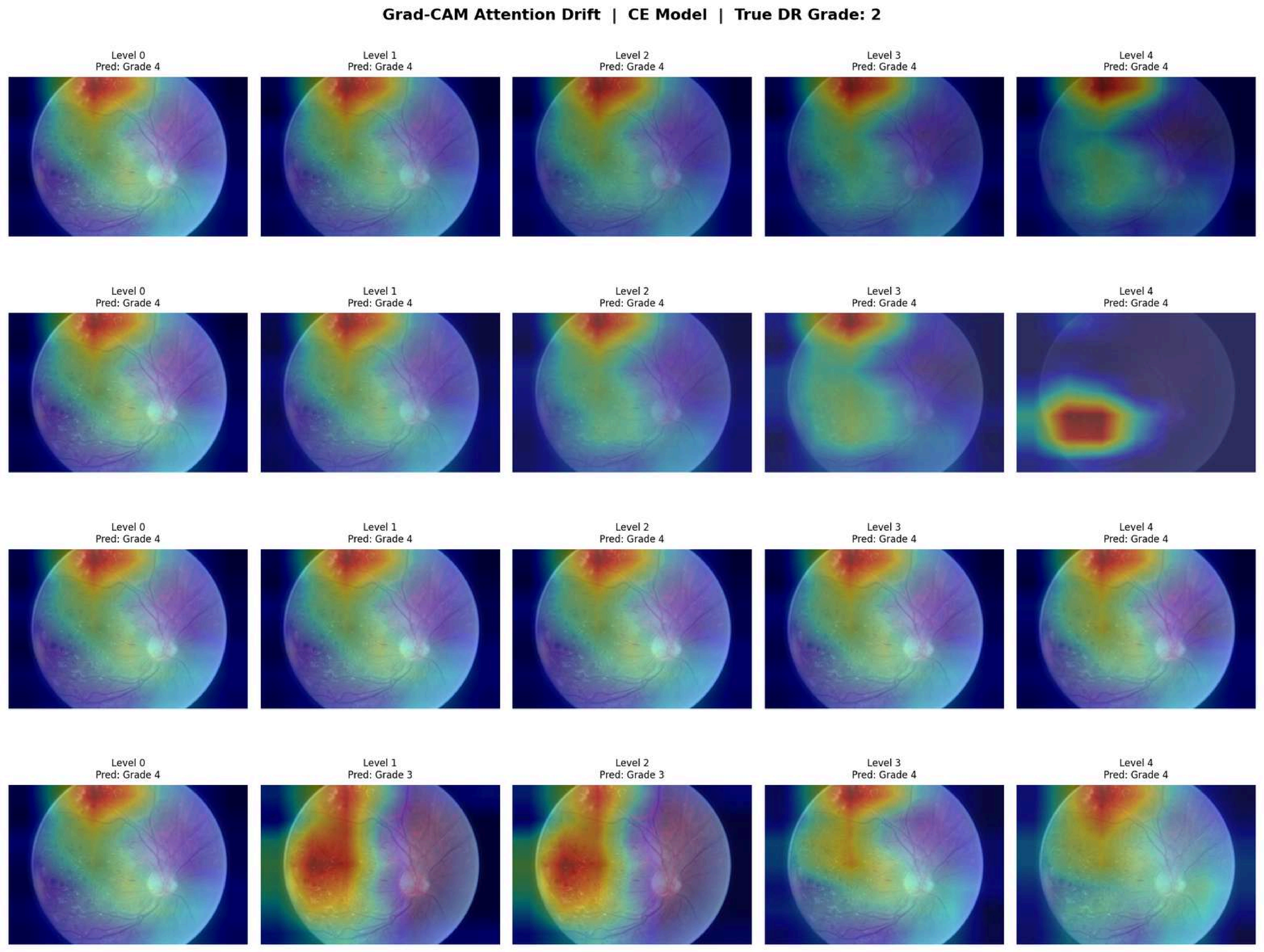
Contrast & General Robustness: The Ordinal model consistently maintains higher QWK scores across brightness and blur, while successfully **delaying the critical performance cliff (QWK < 0.6) under severe contrast degradation until level 4 (~0.32) compared to CE's premature collapse at level 3 (~0.59).**

Noise Vulnerability: Conversely, under maximum noise (severity 4), **the Ordinal model undergoes a late performance cliff, dipping below the threshold to ~0.58, whereas the CE model demonstrates an isolated advantage by holding steady just above the threshold at ~0.62.**

GradCAM Attention Analysis



GradCAM Attention Analysis



Why Ordinal Modeling Wins

Smooth convergence with a 71% reduction in final validation loss **(~0.30 vs 1.05 for CE).**

LIMITATIONS

- **Performance drops under low contrast / poor image quality**, as lesion visibility is reduced and important features may be missed.
- **Training data lacks sufficient heterogeneity**, so the model may not generalize well to diverse real-world clinical conditions.
- **Datasets used (APTOS, IDRiD) are relatively homogeneous**, with similar imaging protocols and populations, limiting exposure to variability across devices, hospitals, and demographics.
- A clear **domain gap exists between training and external datasets** (e.g., Messidor-2), which negatively impacts generalization performance.
- The model **shows confusion between adjacent DR grades**, especially in early stages where visual differences are subtle.
- Limited interpretability and scope: **Grad-CAM provides only partial explanations**, and the model performs only DR grading without lesion detection or multi-disease screening.

CHALLENGES

Computational Challenges

- Limited GPU Resources: Unreliable Kaggle sessions and limited hardware made training large models difficult, requiring optimization of batch size and local execution.

Modeling Challenges

- Loss Function Selection: Required multiple experiments to choose between cross-entropy and ordinal loss; CORAL performed better for DR grading.
- Class Imbalance: Skewed distribution of DR grades biased learning, requiring weighted sampling and balanced loss strategies.
- Hyperparameter Sensitivity: Model performance was highly sensitive to learning rate and augmentation, affecting convergence stability.

Data Challenges

- Data Heterogeneity & Domain Shift: Variations in image quality and differences across datasets (APTOS/EyePACS vs IDRiD, Messidor-2) reduced generalization.
- Limited Interpretability: Difficult to ensure clinically reliable explanations; Grad-CAM provides only partial insights.

Deployability

Can this solution be deployed at Plaksha?

- Our solution focuses on AI-based Diabetic Retinopathy (DR) detection using fundus images
- Currently NOT directly deployable at Plaksha.
- Reason: No fundus imaging equipment (retinal cameras) available on campus
- Model requires retinal images as input, which cannot be captured here.

Where can it be deployed?

- Can be deployed in any hospital or eye clinic that has fundus imaging equipment
- The model can assist doctors in early screening and diagnosis