Erciyes University Journal of Institue Of Science and Technology

Research Article
10.65520/erciyesfen.1970251

Imprint:
Volume: 42(3)
Year: 2026
Page: 187-200

Fatih ŞAHİN[a*]

[a] Asst. Prof., İstanbul Topkapı University, fatihsahin@topkapi.edu.tr

[*] Corresponding Author

Received: 6/13/2026
Accepted: 9/1/2026

Citation:
Fatih ŞAHİN (2026). A Fair Evaluation of Lightweight CNNs and Ordinal Regression for Diabetic Retinopathy Grading: The Double-Rebalancing Pitfall. *Erciyes University Journal of Institue Of Science and Technology*, 42(3), 187-200. https://doi.org/10.65520/erciyesfen.1970251

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

A Fair Evaluation of Lightweight CNNs and Ordinal Regression for Diabetic Retinopathy Grading: The Double-Rebalancing Pitfall

Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness worldwide, affecting approximately one-third of diabetic patients. Early detection through automated retinal screening systems can significantly reduce the burden of vision loss. In this study, we present a comparative analysis of three lightweight convolutional neural networks—EfficientNet-B0, MobileNetV3-Small, and DenseNet-121—for five-class DR severity grading using the APTOS 2019 dataset. We employ a two-phase transfer learning strategy with ImageNet-pretrained backbones, enhanced by CLAHE and Ben Graham preprocessing. To examine how the class-balancing configuration and the loss function affect ordinal DR grading, particularly for Moderate DR detection, we conduct ablation studies comparing CrossEntropy, Focal Loss, and CORN (Conditional Ordinal Regression) loss functions. Our central finding is methodological: combining a class-balanced sampler with an inverse-frequency weighted loss ("double rebalancing") silently collapses the second-largest class (Moderate DR) to an F1-score of 0.00, whereas removing either component restores Moderate-DR F1 to 0.63–0.75. Under a fair single-rebalancing configuration, well-configured lightweight CNNs are strong and generalizable, with the best (CrossEntropy) configuration reaching APTOS QWK ≈ 0.88 and, zero-shot on IDRiD, QWK ≈ 0.72 with referable-DR AUC ≈ 0.95 (CORN slightly lower). Ordinal (CORN) and standard (CE) losses perform comparably and are statistically indistinguishable on APTOS, although on IDRiD CrossEntropy is significantly better on mean absolute error; CORN's practical advantage is robustness, reaching comparable performance without class-weight tuning and never collapsing a class. Grad-CAM is used to qualitatively illustrate model attention. These results indicate that correct class-balancing configuration matters more than the choice of loss function for ordinal DR grading.

Keywords: Diabetic retinopathy, Deep learning, Ordinal regression, CORN loss, Grad-CAM

Diyabetik retinopati derecelendirmesi için hafif CNN'lerin ve sıralı regresyonun adil değerlendirmesi: çifte-dengeleme tuzağı

Öz

Diyabetik retinopati (DR), dünya genelinde önlenebilir körlüğün önde gelen nedenlerinden biridir ve diyabet hastalarının yaklaşık üçte birini etkilemektedir. Bu çalışmada, APTOS 2019 veri kümesi kullanılarak beş sınıflı DR şiddet derecelendirmesi için üç hafif evrişimsel sinir ağının (EfficientNet-B0, MobileNetV3-Small ve DenseNet-121) karşılaştırmalı analizi sunulmaktadır. ImageNet ön eğitimli omurgalarla iki aşamalı transfer öğrenme, CLAHE ve Ben Graham ön işleme ile desteklenmiştir. Sıralı sınıf sınırı

belirsizliğini ele almak için CrossEntropy, Focal Loss ve CORN (Koşullu Sıralı Regresyon) kayıp fonksiyonları karşılaştırılmıştır. Temel bulgumuz metodolojiktir: sınıf-dengeli örnekleyici (sampler) ile ters-frekans ağırlıklı kaybın birlikte kullanımı ("çifte dengeleme"), ikinci en büyük sınıfı (Orta DR) sessizce F1=0.00'a çökertmekte; bileşenlerden biri kaldırıldığında Orta DR F1'i 0.63–0.75'e geri dönmektedir. Adil, tek-dengelemeli yapılandırmada doğru kurulmuş hafif CNN'ler güçlü ve genellenebilirdir; en iyi (CrossEntropy) yapılandırması APTOS QWK ≈ 0.88 ve sıfır-atış IDRiD QWK ≈ 0.72, refere-edilebilir DR AUC ≈ 0.95'e ulaşır (CORN biraz daha düşüktür). Sıralı (CORN) ve standart (CE) kayıplar APTOS'ta benzer performans göstermekte ve istatistiksel olarak ayırt edilememektedir; ancak IDRiD'de ortalama mutlak hata bakımından CrossEntropy anlamlı biçimde daha iyidir. CORN'un pratik avantajı dayanıklılıktır: sınıf-ağırlığı ayarı gerektirmeden benzer performansa ulaşır ve hiçbir sınıfı çökertmez. Grad-CAM, model dikkatini niteliksel olarak göstermek için kullanılmıştır. Bu sonuçlar, sıralı DR derecelendirmesinde doğru sınıf-dengeleme yapılandırmasının kayıp fonksiyonu seçiminden daha belirleyici olduğunu göstermektedir.

Anahtar kelimeler: Diyabetik retinopati, Derin öğrenme, Sıralı regresyon, CORN kaybı, Grad-CAM

1.INTRODUCTION

Diabetes mellitus is a global health crisis affecting over 530 million adults worldwide, with projections estimating 643 million by 2030 [1]. Turkey ranks among the top European countries in diabetes prevalence, with approximately 7 million diagnosed patients and an estimated 15% prevalence rate in adults over 20 years of age [2]. Diabetic retinopathy, a microvascular complication of diabetes, affects approximately 35% of diabetic patients and remains the leading cause of blindness among the working-age population in developed countries [3].

The International Clinical Diabetic Retinopathy (ICDR) severity scale classifies DR into five grades: No DR (Grade 0), Mild NPDR (Grade 1), Moderate NPDR (Grade 2), Severe NPDR (Grade 3), and Proliferative DR (Grade 4). Accurate grading is essential for determining treatment pathways, yet manual grading by ophthalmologists is time-consuming, subjective, and not scalable.

A critical yet underexplored challenge in automated DR grading is the ordinal nature of disease severity. Standard classification approaches using categorical cross-entropy loss treat all misclassifications equally, ignoring the clinical reality that confusing adjacent grades (e.g., Mild vs. Moderate) is far less consequential than confusing distant grades (e.g., No DR vs. Proliferative). This often leads to poor detection of intermediate grades, particularly Moderate DR, which represents a crucial clinical decision boundary for referral.

In this study, we investigate three lightweight CNN architectures—EfficientNet-B0 (~5.3M parameters), MobileNetV3-Small (~2.5M parameters), and DenseNet-121 (~8.0M parameters)—and conduct systematic ablation studies on loss function design. We show that the common practice of combining a class-balanced sampler with an inverse-frequency weighted loss silently collapses Moderate-DR detection to F1=0.00, and that removing either component—or using an ordinal (CORN) loss—restores it; under a fair single-rebalancing configuration, CrossEntropy and CORN perform comparably. We further validate generalizability through cross-dataset evaluation on IDRiD and provide Grad-CAM explainability analysis, as the dataset class distributions are summarized in Figure 1.

CC: BY-NC-ND

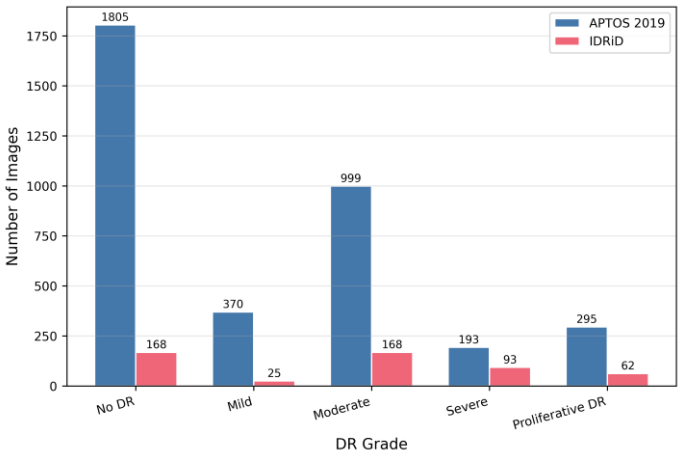

Figure 1. Dataset class distribution for APTOS 2019 and IDRiD.

2. RELATED WORK

2.1. CNN-based diabetic retinopathy classification

Deep learning for DR screening has evolved rapidly since Gulshan et al. [4] demonstrated ophthalmologist-level detection using Inception-v3. Subsequent CNN-based approaches include lesion-focused microaneurysm detectors [5], cross-disease attention networks [6], and high-performing multi-model ensembles [7], [8], [9]. Table 1 summarizes representative results on APTOS 2019.

Table 1. Comparison results on APTOS 2019 five-class DR grading.

Study	Architecture	Params	QWK	Accuracy	Key technique
Kaggle 1st (2019)	EfficientNet-B3 ×15	~180M	0.936	—	Heavy ensemble
Bodapati et al. (2020)	VGG-19	~144M	0.87	0.82	CLAHE + augmentation
Qummar et al. (2019)	5-model ensemble	~200M+	0.85	0.80	VGG+ResNet+Dense
Sikder et al. (2021)	Ensemble (4 CNN)	~100M+	—	0.80	Severity classification
Ours (fair CE/CORN)	DenseNet-121	~8M	0.88	0.790	Ordinal regression

As summarized in Table 1, prior high-performing systems achieve strong agreement largely through scale: the Kaggle-winning solution and the ensembles of Qummar et al. [7], Sikder et al. [8] and Bodapati et al. [9] reach QWK 0.80–0.94 but rely on very large or multi-model architectures (roughly 100–200M+ parameters), which limits deployability on resource-constrained hardware; lighter single-network approaches such as Pratt et al. [5] and the CANet attention model [6] are more efficient but report weaker five-class ordinal agreement. A limitation common to these works is that they optimize standard categorical objectives and do not explicitly model the ordinal structure of DR severity. Our work differs in two key aspects: (1) we exclusively use lightweight models (≤8M parameters), making our approach 8–25× more parameter-efficient than the ensemble methods above; and (2) we systematically evaluate ordinal-aware loss functions, showing that the class-balancing configuration, rather than the loss function, is the decisive factor for five-class DR grading. To our knowledge, this combination of parameter efficiency and an ordinal-aware objective is not addressed by the prior studies in Table 1.

2.2. Transfer learning for medical imaging

Transfer learning from ImageNet-pretrained models is the dominant paradigm in medical image analysis, and more capable ImageNet backbones tend to transfer better to downstream tasks [10]. Raghu et al. [11] demonstrated consistent advantages over training from scratch, while two-phase fine-tuning—frozen classifier training followed by full network fine-tuning—improves convergence [12]. We adopt this established strategy across all experiments.

2.3. Ordinal regression in medical imaging

DR severity follows an inherent ordinal progression, making ordinal regression a natural framework. Shi et al. [13] proposed CORN (Conditional Ordinal Regression for Neural Networks), which decomposes K-class ordinal classification into K-1 conditional binary tasks, each predicting $P(Y > k \mid Y \geq k)$. This formulation guarantees rank consistency without requiring specialized architectures and has shown strong performance for ordinal grading tasks in medical imaging.

2.4. Class imbalance and Focal Loss

Lin et al. [14] introduced Focal Loss to address class imbalance by down-weighting well-classified examples with a focusing parameter γ. In DR classification, the dominance of "No DR" samples creates severe imbalance. Focal Loss has shown promise in improving minority class detection, though its effectiveness for ordinal tasks with ambiguous boundaries remains limited compared to ordinal-specific approaches.

2.5. Explainability in medical AI

Grad-CAM [15] is the standard for CNN attention visualization in retinal images, enabling clinicians to verify predictions are based on pathologically relevant regions. Clinical adoption of AI-based DR screening requires such interpretability guarantees [16].

3. METHODOLOGY

3.1. Image preprocessing

Raw retinal fundus images undergo a multi-stage pipeline: (1) circular cropping to remove black borders; (2) CLAHE on the L channel in LAB color space (clipLimit=2.0, tileGridSize=8×8); (3) Ben Graham preprocessing (addWeighted with Gaussian blur subtraction, σ=10); (4) resize to 224×224; (5) ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). The full pipeline is illustrated in Figure 2.

Figure 2. Preprocessing pipeline: original image, cropped fundus, CLAHE enhancement, Ben Graham normalization, and final 224 × 224 output.

3.2. Data augmentation

Training augmentation includes random horizontal/vertical flips (p=0.5), rotation (±30°), color jitter (brightness=0.2, contrast=0.2, saturation=0.1), and affine transforms (translate=5%, scale=0.9–1.1). Class imbalance is addressed via weighted random sampling.

3.3. Model architectures

Three lightweight architectures with ImageNet pretrained weights are evaluated.

EfficientNet-B0 (~5.3M params): compound-scaled [17], Dropout(0.3) → Linear(1280, 5).

MobileNetV3-Small (~2.5M params): NAS-optimized [18], Linear(576, 256) → Hardswish → Dropout(0.3) → Linear(256, 5).

CC: BY-NC-ND

DenseNet-121 (~8.0M params): dense connectivity [19], Linear(1024, 5).

3.4. Training strategy

We use a two-phase transfer learning schedule. Phase 1 (Epochs 1–5) trains the classifier only with a frozen backbone (AdamW, lr=1e-4). Phase 2 (Epochs 6–30) performs full fine-tuning (lr=1e-5, CosineAnnealingLR, T_max=25). Early stopping uses patience=7 while monitoring validation QWK.

3.5. Loss function ablation

Three loss functions are systematically compared. CrossEntropy (CE) is the standard categorical loss. Unless otherwise stated, the baseline training pipeline applies two class-balancing mechanisms simultaneously—a class-balanced WeightedRandomSampler and inverse-frequency class weights in the loss; the effect of this "double rebalancing" is examined by ablation in Section 5 (Table 5). Focal Loss (FL), defined in Equation (1), uses γ=2.0 and class-balanced α to down-weight easy examples and focus on hard boundary cases.

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (1)$$

CORN (Conditional Ordinal Regression) decomposes the 5-class ordinal task into 4 conditional binary classifiers, each predicting $P(Y > k \mid Y \geq k)$. The classifier head outputs 4 logits instead of 5 class scores, and class probabilities are recovered via the chain rule of conditional probability. This formulation naturally respects the ordinal structure of DR grades.

3.6 Grad-CAM explainability

Grad-CAM visualizations from the final convolutional layers of each model (3 samples per class × 3 models = 45 heatmaps) identify regions driving classification decisions.

4. EXPERIMENTAL SETUP

4.1. Datasets

APTOS 2019: 3,662 retinal fundus images (train=2,930, val=366, test=366) with five-class ICDR labels. Class distribution: No DR=49.3%, Mild=10.1%, Moderate=27.3%, Severe=5.3%, Proliferative=8.1%.

IDRiD: 516 retinal images (train=413, test=103) with five-class ICDR grading for cross-dataset validation. Notable distribution shift: the Severe class is 3.4× more prevalent in IDRiD (18.0% vs 5.3%), while No DR is underrepresented (32.6% vs 49.3%).

4.2. Evaluation metrics

The primary metric is Quadratic Weighted Kappa (QWK), which accounts for ordinal misclassification costs. Secondary metrics are accuracy and F1-score (macro/weighted) with per-class precision/recall. Ordinal behaviour is measured via adjacency accuracy (±1 and ±2 grade tolerance), supported by normalized confusion matrices and per-class ROC/AUC. AUC-macro is computed as the unweighted mean of the five one-vs-rest per-class AUCs (scikit-learn roc_curve/auc). Statistical analysis uses bootstrap 95% confidence intervals (1000 iterations) and McNemar's test for pairwise model comparison.

4.3. Implementation details

Experiments use PyTorch 2.11.0 (torchvision 0.26.0, CUDA 13.0) with Python 3.10.19, NumPy 2.2.6, and scikit-learn 1.7.2, on an NVIDIA RTX 5090 GPU (32 GB), batch size=32, 4 workers, seed=42 (with torch/NumPy/random seeded and cudnn.deterministic=True). Training time per model is approximately 30–42 minutes.

CC: BY-NC-ND

5. RESULTS

5.1. Baseline model comparison (CrossEntropy)

Baseline performance on the APTOS test set (CE loss, n=366) is reported in Table 2.

Table 2. Baseline performance on APTOS test set (CE loss, n=366).

Model	QWK [95% CI]	Accuracy	Adj Acc ±1	F1-macro	AUC
EfficientNet-B0	0.815 [0.781, 0.849]	0.639	—	0.406	0.902
MobileNetV3-Small	0.778 [0.738, 0.818]	0.607	—	0.384	0.891
DenseNet-121	0.816 [0.780, 0.851]	0.626	—	0.417	0.899

A McNemar test for EfficientNet-B0 vs DenseNet-121 showed no statistically significant difference ($p > 0.05$), confirming comparable baseline performance. A critical finding is that all baseline models completely failed to detect Moderate DR (F1 = 0.00 across all three architectures), despite Moderate being the second-largest class (27.3% of training data). Training dynamics are shown in Figure 3.

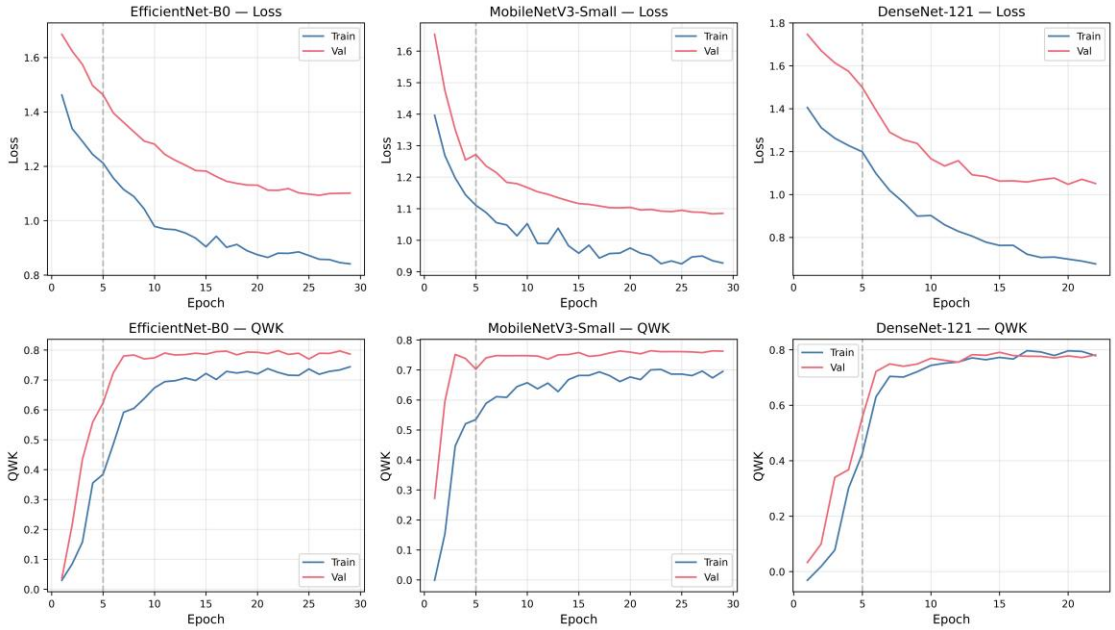

Figure 3. Training and validation loss and QWK curves for EfficientNet-B0, MobileNetV3-Small, and DenseNet-121.

5.2. Loss function ablation study

Table 3 reports DenseNet-121 performance by loss function on the APTOS test set.

Table 3. DenseNet-121 performance by loss function on APTOS test set.

Loss function	QWK [95% CI]	Accuracy	Adj Acc ±1	F1-macro	AUC
CrossEntropy	0.816 [0.780, 0.851]	0.626	—	0.417	0.899
Focal Loss (γ=2)	0.793 [0.751, 0.832]	0.587	0.902	0.411	0.893
CORN	0.875 [0.834, 0.911]	0.790	0.962	0.608	0.924

Relative to the default double-rebalanced CrossEntropy baseline, CORN appears to yield large gains (for example +5.9 QWK points); the ablation in Table 5, however, shows that these apparent gains are largely eliminated once the CrossEntropy baseline is configured fairly, so they reflect the mis-configured baseline rather than an intrinsic advantage of the ordinal loss.

CC: BY-NC-ND

Table 4. Per-class F1-score comparison (DenseNet-121).

Class	Loss	Precision	Recall	F1
No DR	CE	0.989	0.889	0.937
	Focal	0.994	0.804	0.889
	CORN	0.975	0.970	0.972
Mild	CE	0.294	0.833	0.435
	Focal	0.247	0.800	0.378
	CORN	0.391	0.600	0.474
Moderate	CE	0.000	0.000	0.000
	Focal	0.857	0.069	0.128
	CORN	0.722	0.655	0.687
Severe	CE	0.207	0.706	0.320
	Focal	0.200	0.647	0.306
	CORN	0.333	0.647	0.440
Prolif. DR	CE	0.349	0.455	0.395
	Focal	0.304	0.424	0.354
	CORN	1.000	0.303	0.465

Under this default double-rebalanced configuration, CORN attains Moderate-DR F1=0.687 (precision 0.722, recall 0.655) where CrossEntropy collapses to 0.00; as the ablation in Table 5 shows, standard CrossEntropy recovers comparable Moderate-DR detection (F1=0.70–0.75) once either rebalancing component is removed. Focal Loss achieved limited success (F1=0.128) with very high precision (0.857) but extremely low recall (0.069), indicating it learned to make very few but accurate Moderate predictions rather than truly solving the classification boundary problem. Baseline metrics are compared in Figure 4, confusion matrices in Figure 5, and the loss ablation per-class F1 in Figure 6.

Table 5. Class-balancing ablation on DenseNet-121 (APTOS test, n=366, seed=42). Only the double-rebalanced CrossEntropy configuration collapses Moderate DR to F1=0.00; removing either the sampler or the inverse-frequency loss weighting restores it. CORN has no class-weighting axis, so its two configurations differ only by the sampler. Note: Tables 3–4 report each model's original training run, whereas this ablation uses a unified re-implementation in which every configuration shares an identical training loop (same optimizer, schedule, and epoch budget); its absolute values therefore differ slightly from Tables 3–4 (e.g., CORN QWK 0.863 vs 0.875), while all comparisons within the ablation are exact and seed-matched (seed=42).

Configuration	Sampler	Wt. loss	Moderate F1	QWK	Acc.	F1-macro
CE (double-rebalanced)	Yes	Yes	0.00	0.816	0.631	0.420
CE (weighted loss only)	No	Yes	0.635	0.866	0.765	0.553
CE (sampler only)	Yes	No	0.703	0.881	0.792	0.593
CE (plain)	No	No	0.751	0.872	0.820	0.549
CORN (sampler)	Yes	—	0.642	0.863	0.773	0.579
CORN (plain)	No	—	0.763	0.861	0.820	0.553

CC: BY-NC-ND

Table 6. Multi-seed reliability under the fair (plain) configuration (DenseNet-121, five seeds; mean ± std, evaluated on APTOS test and zero-shot on IDRiD). CrossEntropy and CORN are comparable.

Configuration	APTOS QWK	APTOS Mod-F1	APTOS F1-macro	IDRiD QWK	IDRiD ref-AUC
CE (plain)	0.881±0.007	0.760±0.009	0.570±0.014	0.707±0.022	0.948±0.003
CORN (plain)	0.860±0.014	0.740±0.018	0.542±0.021	0.658±0.025	0.944±0.001

Table 7. Paired-bootstrap comparison of CORN against a fair CE baseline (CE with sampler only; matched pipeline; 10,000 resamples). No metric significantly favors CORN.

Dataset	Metric	Δ (CORN−CE)	95% CI	Significant?
APTOS	QWK	−0.017	[−0.045, +0.011]	No
APTOS	Adjacency ±1	0.000	[−0.019, +0.019]	No
APTOS	MAE (grade)	+0.022	[−0.027, +0.071]	No
IDRiD	QWK	−0.042	[−0.086, +0.004]	No
IDRiD	Adjacency ±1	+0.029	[−0.002, +0.060]	No
IDRiD	MAE (grade)	+0.099	[+0.033, +0.163]	Yes (favors CE)

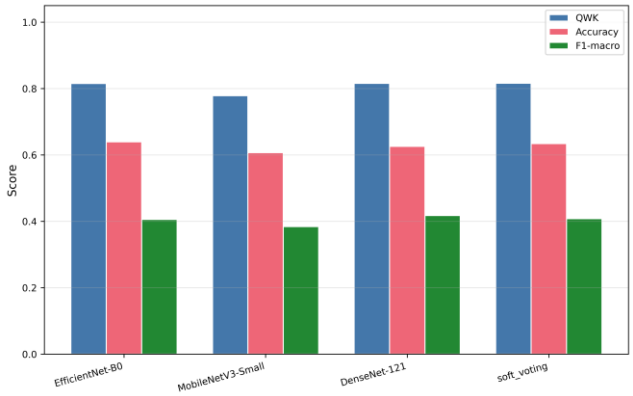

Figure 4. Baseline model comparison on APTOS using QWK, accuracy, and F1-macro.

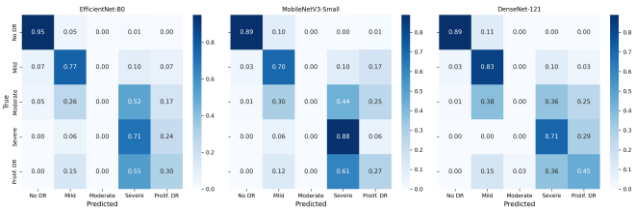

Figure 5. Normalized confusion matrices for EfficientNet-B0, MobileNetV3-Small, and DenseNet-121.

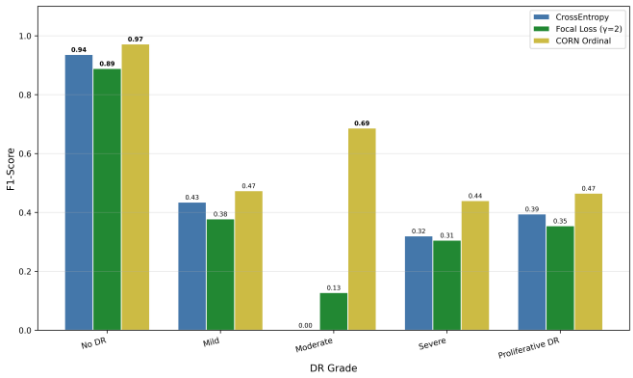

Figure 6. Loss function ablation on DenseNet-121 showing per-class F1 for CrossEntropy, Focal Loss, and CORN.

5.3 Grad-CAM analysis

Grad-CAM visualizations provide qualitative interpretability, illustrating which retinal regions influence each model's predictions rather than evidence of clinically correct attention, as no lesion-level annotations or expert-reader validation were used. Qualitatively, EfficientNet-B0 tends to focus on microaneurysm- and hemorrhage-rich regions in early grades, DenseNet-121 shows broader attention over vascular abnormalities in severe cases, and for Proliferative DR the models tend to highlight regions near the optic disc and retinal periphery (Figure 7).

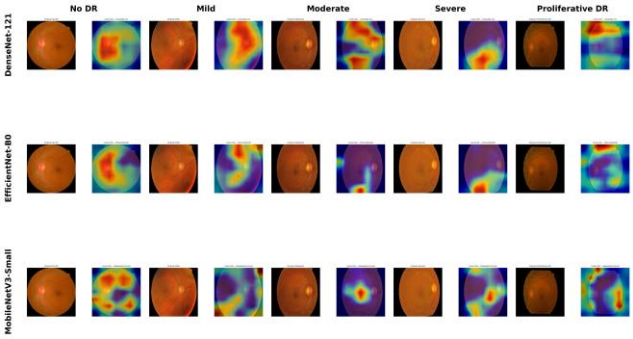

Figure 7. Grad-CAM visualizations across the five DR grades for DenseNet-121, EfficientNet-B0, and MobileNetV3-Small.

5.4 Cross-dataset validation

Cross-dataset performance of APTOS-trained models on IDRiD (n=516) is reported in Table 8.

Table 8. Cross-dataset performance: APTOS-trained models on IDRiD (n=516).

Model	QWK [95% CI]	Accuracy	Adj Acc ±1	Adj Acc ±2	AUC
EfficientNet-B0	0.626 [0.579, 0.674]	0.223	0.841	0.971	0.758
MobileNetV3-Small	0.568 [0.510, 0.623]	0.200	0.835	0.957	0.756
DenseNet-121	0.606 [0.558, 0.648]	0.190	0.785	0.969	0.794

Under fair configurations, both CrossEntropy and CORN transfer to IDRiD without any retraining (Table 9): fair CrossEntropy reaches QWK ≈ 0.70–0.72 and CORN ≈ 0.68–0.70—comparable cross-dataset performance—while CORN retains the highest ±1-grade adjacency. For clinically oriented referable-DR detection (Table 10), fair CrossEntropy matches or exceeds CORN on AUC and sensitivity, whereas CORN attains the highest specificity.

Table 9. Zero-shot cross-dataset performance on IDRiD (n=516), DenseNet-121, fair configurations. Under fair baselines CORN does not outperform CrossEntropy on QWK; its residual strength is the highest ±1-grade adjacency.

Configuration	QWK	Acc.	Adj ±1	F1-macro	AUC-macro
CE (sampler only)	0.724	0.537	0.847	0.416	0.812
CE (plain)	0.697	0.576	0.820	0.371	0.806
CORN (sampler)	0.682	0.417	0.876	0.364	0.801
CORN (plain)	0.697	0.568	0.818	0.355	0.791

Table 10. Referable-DR (grades ≥ 2) detection on IDRiD (zero-shot), DenseNet-121, fair configurations. Fair CrossEntropy matches or exceeds CORN on AUC and sensitivity, while CORN attains the highest specificity.

Configuration	AUC	Sensitivity	Specificity
CE (sampler only)	0.944	0.879	0.902
CE (plain)	0.949	0.873	0.922
CORN (sampler)	0.923	0.786	0.953
CORN (plain)	0.942	0.867	0.917

The conventional accuracy metric (~20%) substantially understates model performance on IDRiD. Adjacency accuracy (±1 grade tolerance) reveals that up to 84% of predictions are within one DR grade of the true label (0.785–0.841 across models), and ±2 tolerance reaches 97%. This demonstrates that models largely preserve ordinal ranking even under domain shift, with errors predominantly occurring between adjacent grades rather than distant ones.

Domain shift analysis shows that IDRiD's distribution differs substantially from APTOS: Severe DR is 3.4× more prevalent (18.0% vs 5.3%), while No DR is underrepresented (32.6% vs 49.3%). Confusion matrix analysis reveals that APTOS-trained models systematically under-predict severe grades on IDRiD, likely because the training distribution biases models toward lower severity predictions. Additionally, differences in imaging equipment (fundus camera models, resolution) and patient demographics (Indian vs international populations) contribute to the domain gap. Per-class ROC curves are shown in Figure 8.

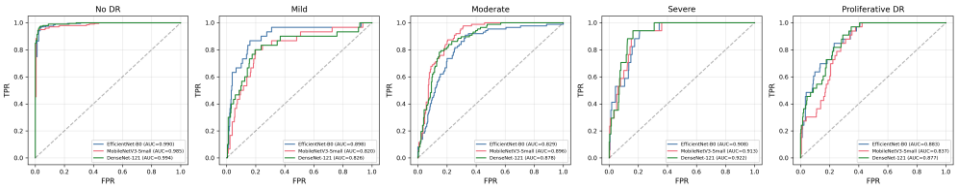

Figure 8. Per-class ROC curves for the three individual models.

A summary comparison of cross-dataset behaviour between APTOS and IDRiD is presented in Figure 9.

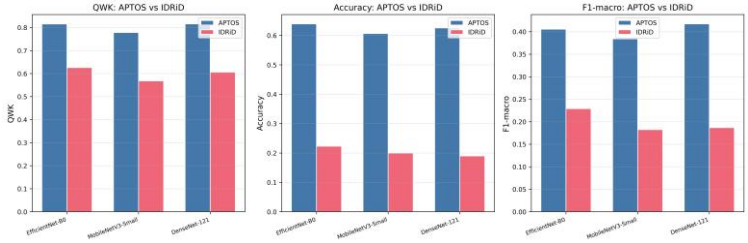

Figure 9. Cross-dataset comparison between APTOS and IDRiD for QWK, accuracy, and F1-macro.

6. DISCUSSION

6.1. The double-rebalancing pitfall

Our most significant finding is methodological. The complete failure of the baseline models to detect Moderate DR (F1=0.00) was initially attributed to ordinal boundary ambiguity, but a controlled ablation shows that this collapse is an artifact of the training configuration rather than a property of the loss function. The baseline combined two rebalancing mechanisms simultaneously—a class-balanced WeightedRandomSampler and an inverse-frequency weighted CrossEntropy loss ("double rebalancing"). Because Moderate is the second-largest class (27.3%), inverse-frequency weighting assigns it a low weight (0.73) while up-weighting its rarer neighbours (Severe 3.81, Proliferative 2.50); combined with a sampler that further oversamples those neighbours, the decision boundary is pushed away from Moderate until the class receives essentially no predictions (F1=0.00). Removing either component restores Moderate-DR F1 to 0.63–0.75 (Table 5): with the sampler alone and unweighted loss Moderate F1=0.70, and with neither 0.75. The collapse is therefore a silent side-effect of aggressive double rebalancing, not evidence that CrossEntropy cannot learn the Mild/Moderate boundary. Under a fair single-rebalancing configuration the choice of loss has little effect: standard CrossEntropy and CORN are statistically indistinguishable on APTOS (paired bootstrap, all 95% CIs include zero). CORN decomposes the task into four rank-consistent conditional binary decisions and, by construction, is unaffected by class weighting, so its only rebalancing axis is the sampler, across which its QWK is essentially invariant (0.863 vs 0.861); CrossEntropy's QWK instead ranges from 0.816 to 0.881 across configurations and, when double-rebalanced, drops an entire class. CORN's practical value is therefore robustness: competitive performance without class-weight tuning and no silent class collapse.

DenseNet-121 was selected as the primary model for the detailed ablation as the best-performing baseline. The Moderate-DR collapse seen under the default double-rebalanced CrossEntropy configuration (Table 3) is a property of the training configuration rather than of the architecture or the loss: as Table 5 shows, removing either rebalancing component restores Moderate-DR F1 to 0.63–0.75, and an ordinal (CORN) loss—which has no class-weighting term—reaches comparable performance. Under a fair configuration the two losses are statistically indistinguishable on APTOS (Table 7).

6.2. Why Focal Loss degraded baseline performance

Under the default double-rebalanced configuration, a counterintuitive finding is that Focal Loss (QWK=0.793) performed worse than standard CrossEntropy (QWK=0.816), despite being specifically designed to address class imbalance. This degradation can be explained by the fundamental mismatch between Focal Loss's assumptions and the nature of ordinal classification boundaries. Focal Loss down-weights "easy" examples (those classified with high confidence) through the $(1-p_t)^{\gamma}$ modulating factor, redirecting gradient signal toward "hard" examples. However, in ordinal grading tasks with ambiguous inter-class boundaries, this easy/hard distinction becomes misleading. Samples near the Mild/Moderate boundary are inherently ambiguous due to the continuous nature of disease progression—they are not "hard" due to insufficient model capacity, but rather due to genuine label uncertainty. By aggressively focusing on these boundary cases, Focal Loss amplifies noisy gradient signals from samples whose ground-truth labels may themselves be inconsistent, ultimately destabilizing the learned decision boundaries. Furthermore, Focal Loss with class-balanced α weights introduces two competing reweighting mechanisms simultaneously—frequency-based (α) and confidence-based (γ)—which can interfere with each other in the presence of ordinal structure. The net effect is a model that sacrifices overall discrimination to marginally improve Moderate DR detection (F1: 0.000→0.128), a poor trade-off compared to CORN's principled ordinal decomposition.

6.3. Lightweight model efficiency

Our CORN-trained DenseNet-121 achieves QWK=0.875 with only ~8M parameters, competitive with published results using models 8–25× larger (Table 1). This demonstrates that appropriate loss function design can compensate for reduced model capacity, a finding with significant implications for deploying DR screening in resource-constrained settings.

CC: BY-NC-ND

6.4. Re-evaluating cross-dataset performance

The large discrepancy between accuracy (~20%) and QWK (~0.62) on IDRiD initially appears alarming. However, adjacency accuracy analysis reveals that 84% of predictions fall within ±1 grade of truth, and 97% within ±2 grades. In clinical practice, ±1 grade tolerance is often considered acceptable for screening purposes, as treatment decisions are primarily driven by the distinction between referable (Grades 2–4) and non-referable (Grades 0–1) DR. The domain shift is driven by three factors: (1) class distribution differences—IDRiD has 3.4× more Severe DR than APTOS; (2) imaging equipment heterogeneity between datasets; (3) annotation protocol differences between crowdsourced (APTOS) and expert-graded (IDRiD) labels.

6.5. Clinical implications

The models' ±1-grade adjacency accuracy of 96.2% on APTOS suggests potential utility for clinical screening workflows where ±1 grade tolerance is acceptable, and its lightweight architecture is compatible with standard hardware. We emphasize, however, that these results are based on retrospective public datasets and do not include prospective validation, expert-reader comparison, or model calibration analysis; such evaluation would be required before any clinical deployment. Grad-CAM is offered as qualitative interpretability rather than a guarantee of clinically correct reasoning.

6.6. Limitations

(1) Cross-dataset validation is limited to IDRiD; additional external datasets (e.g., Messidor-2, EyePACS) would further strengthen generalizability claims. (2) Per-class sample sizes in the APTOS test set are small (Severe: n=17, Proliferative: n=33), contributing to wider confidence intervals for these classes. (3) This is a retrospective study on public datasets and does not include prospective validation, expert-reader comparison, or model calibration analysis, which would be required before clinical deployment. (4) Grad-CAM provides qualitative interpretability only; lesion-level or expert-validated attention assessment remains future work.

7. CONCLUSION

This study presents a fair comparison of lightweight CNNs for five-class diabetic retinopathy grading and identifies a methodological pitfall relevant to ordinal medical-image classification. Combining a class-balanced sampler with an inverse-frequency weighted loss ("double rebalancing") silently collapses Moderate-DR F1 to 0.00; removing either component restores it to 0.63–0.75. This collapse is a configuration artifact, not evidence that ordinal-aware loss is required. Under a fair single-rebalancing configuration, standard CrossEntropy and CORN are statistically indistinguishable on APTOS (QWK 0.86–0.88, F1-macro ≈ 0.55–0.59); on IDRiD the two remain comparable, except that CrossEntropy is significantly better on mean absolute error. The best (CrossEntropy) configuration generalizes zero-shot to IDRiD at QWK ≈ 0.72 and referable-DR AUC ≈ 0.95 (sensitivity ≈ 0.88), with CORN slightly lower. Lightweight models (≤8M parameters) thus achieve competitive, deployable performance provided the class-balancing configuration is correct. Adjacency (±1-grade) accuracy of up to ~85% on IDRiD indicates screening potential despite lower exact accuracy under domain shift. Overall, for ordinal DR grading the class-balancing configuration matters more than the choice of loss function; CORN's practical contribution is robustness—competitive accuracy without class-weight tuning and no silent class collapse.

8. FUTURE WORK

Future directions include: (1) extending CORN training to all architectures and building CORN-based ensembles; (2) domain adaptation techniques for cross-dataset robustness; (3) incorporating lightweight vision transformers (MobileViT); (4) a prospective clinical validation study; and (5) joint DR and diabetic macular edema detection.

CC: BY-NC-ND

Peer-review: External, Independent.

Acknowledgements:

-

Declarations:

1. Statement of Originality:

The author declares that this manuscript is original, has not been published previously, and is not under consideration for publication elsewhere.

2. Author Contributions:

Concept: FŞ; ***Conceptualization:*** FŞ; ***Literature Search:*** FŞ; ***Data Collection:*** FŞ; ***Data Processing:*** FŞ; ***Analysis:*** FŞ; ***Writing – original draft:*** FŞ; ***Writing – review & editing:*** FŞ.

3. Ethics approval:

This study used publicly available and fully de-identified retinal image datasets (APTOS 2019 and IDRiD); therefore, ethical approval is not required for this study.

4. Funding/Support:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

5. Competing Interests:

The author declares no competing interests.

6. GenAI Usage Statement:

Generative AI tools were used solely for English-language translation and spelling/grammar checking; all scientific content, analyses, and conclusions are the author's own.

7. Sustainable Development Goals:

REFERENCES

1. Magliano, D. J., Boyko, E. J., IDF Diabetes Atlas 10th Edition Scientific Committee. 2021. IDF Diabetes Atlas. 10th ed. International Diabetes Federation, Brussels.
2. Satman, I., Omer, B., Tutuncu, Y., Kalaca, S., Gedik, S., Dinccag, N. et al. 2013. Twelve-year trends in the prevalence and risk factors of diabetes and prediabetes in Turkish adults. European Journal of Epidemiology, 28(2), 169-180.
3. Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T. et al. 2012. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care, 35(3), 556-564.
4. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A. et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.
5. Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., Zheng, Y. 2016. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science, 90, 200-205.

6. Li, X., Hu, X., Yu, L., Zhu, L., Fu, C.-W., Heng, P.-A. 2019. CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Transactions on Medical Imaging, 39(5), 1483-1493.

7. Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., Khan, W. 2019. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access, 7, 150530-150539.

8. Sikder, N., Masud, M., Bairagi, A. K., Arif, A. S. M., Nahid, A.-A., Alhumyani, H. A. 2021. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry, 13(4), 670.

9. Bodapati, J. D., Naralasetti, V., Shareef, S. N., Hakak, S., Bilal, M., Maddikunta, P. K. R., Jo, O. 2020. Blended multi-modal deep ConvNet features for diabetic retinopathy severity prediction. Electronics, 9(6), 914.

10. Kornblith, S., Shlens, J., Le, Q. V. 2019. Do better ImageNet models transfer better? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2661-2671.

11. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S. 2019. Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems, 32.

12. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., Liang, J. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging, 35(5), 1299-1312.

13. Shi, X., Cao, W., Raschka, S. 2023. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Analysis and Applications, 26(3), 941-955.

14. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P. 2017. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.

15. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626.

16. Quellec, G., Charriere, K., Boudi, Y., Cochener, B., Lamard, M. 2017. Deep image mining for diabetic retinopathy screening. Medical Image Analysis, 39, 178-193.

17. Tan, M., Le, Q. V. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 6105-6114.

18. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M. et al. 2019. Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1314-1324.

19. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q. 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700-4708.

CC: BY-NC-ND