Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 25 May 2026
Revised 29 June 2026
Accepted 17 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 1324–1344

KEYWORDS:
Diabetic retinopathy; reproducibility; seed variance; model comparison; cross-dataset evaluation; medical image classification.

Training Seed Variance Exceeds Architectural Differences in Lightweight Diabetic Retinopathy Grading

Himansu Sekhar Rout[1]*, Saravana Kumar S[2]

[1]Research Scholar, Presidency School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
[2]Associate Professor, Presidency School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
*Corresponding author: himansu.20233cse0035@presidencyuniversity.in

ABSTRACT: Architectural choices in automated diabetic retinopathy grading are routinely justified by a single training run on a single benchmark. This paper asks whether the differences such comparisons report are large enough to be interpretable. Four variants of a compact grading network, differing only in how encoder features are aggregated, together with an EfficientNet-B1 baseline, were each trained five times from different random seeds under otherwise identical conditions on APTOS 2019, giving twenty-five runs, and every run was evaluated both on a frozen APTOS test partition and, with weights unchanged, on the independent IDRiD dataset. Between-architecture differences are smaller than the variation between repeated runs of the same architecture. The spread between architecture means is 0.012 quadratic weighted kappa (QWK) on APTOS and 0.013 on IDRiD, against median seed-to-seed standard deviations of 0.010 and 0.027 respectively; on the external dataset the noise is twice the signal. Four of the five architectures achieve the best score on at least one seed on each dataset, and most span the full range from first to last. A single-seed experiment conducted first appeared to show the ranking reversing between the two datasets, with an architecture-by-dataset interaction of +0.052 QWK ($p = 0.012$); across five seeds that interaction has mean −0.006 and standard deviation 0.039, and no model pair shows a stable interaction. Findings that do survive replication are reported: all models compress predictions toward the middle of the ordinal scale under transfer, mild disease collapses from an F1 of 0.46–0.60 to 0.14–0.18, and ImageNet pretraining separates from scratch training by roughly four times the observed seed variance. The practical conclusion is that architectural rankings in this regime require repeated training to be meaningful, and that single-run comparisons, including those in recent work, should be read as provisional.

1. INTRODUCTION

Diabetic retinopathy (DR) is a leading cause of preventable blindness among working-age adults, and its progression is graded on the five-point International Clinical Diabetic Retinopathy (ICDR) severity scale, ranging from no retinopathy to proliferative disease (Wilkinson et al., 2003). Screening programmes generate far more fundus photographs than trained readers can assess, which has motivated a large body of work on automated grading, beginning with large-scale demonstrations of physician-level referral accuracy (Gulshan et al., 2016).

Two properties of the task shape how such systems should be built and evaluated. First, the grades are ordinal: mistaking moderate for severe disease is a far smaller error than mistaking healthy for proliferative, and this asymmetry is reflected in quadratic weighted kappa (QWK) (Cohen, 1968), the metric adopted by the APTOS 2019 challenge and used throughout this paper. Second, the evidence is sparse and small in scale. Grade one is defined by the presence of microaneurysms only, which occupy a few pixels in an image of several megapixels, so a grading decision depends on a tiny fraction of the visual field.

Most published DR systems are trained and evaluated on a single dataset with a single held-out split. This practice cannot distinguish a model that recognises lesions from one that has learned dataset-specific colour balance, field-of-view geometry or class priors. Since screening deployments necessarily encounter cameras and populations different from those in the training data, the distinction matters.

This paper makes three contributions.

A measurement of seed variance against architectural difference. Five architectures are each trained five times under identical conditions and evaluated on two datasets. The spread between architecture means is comparable to the seed-to-seed spread of a single architecture on APTOS and smaller than it on IDRiD. Four of the five architectures rank first on at least one seed on each dataset. Architectural comparisons of this size, reported from one run, are therefore not interpretable.

A worked example of how a single-seed result misleads. An earlier single-seed experiment appeared to show the ranking of aggregation variants reversing between the two datasets, with an interaction of +0.052 QWK and $p = 0.012$ under a difference-of-differences bootstrap. Repeating the experiment over five seeds gives a mean interaction of −0.006 with standard deviation 0.039; the full reversal occurs in one seed of five. The original result is reported here alongside its refutation, because the sequence illustrates how a well-constructed significance test on a single run can still describe noise.

Findings that survive replication. Degradation under transfer is systematic rather than diffuse: every model compresses predictions toward the middle of the ordinal scale, under-assigns the sight-threatening grades, and collapses on mild disease, whose F1 falls from 0.46–0.60 within APTOS to 0.14–0.18 on IDRiD. ImageNet pretraining separates from-scratch training by roughly four times the observed seed variance, making it the one architectural factor examined here whose effect exceeds training noise.

The experimental vehicle is CARETLite, a family of compact grading networks pairing a pretrained EfficientNet-B0 encoder with a lightweight aggregation head. It is deliberately simple: it exists to make the aggregation stage the only variable across conditions, so that any difference between variants is attributable to that stage. Its absolute performance is not the subject of this paper, and the variants are not proposed as an advance on the state of the art.

The motivation for repeating each configuration was to assess the stability of the observed results across repeated evaluations. This additional analysis changed the interpretation of the findings and provided stronger evidence for the conclusions drawn from the experiments.

2. RELATED WORK

2.1 Architectural design for DR grading

Deep convolutional networks have dominated automated DR grading since large annotated collections became available, and recent architectures pursue accuracy by combining local and global feature modelling. Yi et al. (2026) propose L-MedViT, a hierarchical hybrid network integrating a Kolmogorov-Arnold network and a lesion-aware module within a transformer, together with a focus-attention mechanism that processes lesions of differing size through multi-scale windows; they report 88.47 percent accuracy on APTOS 2019 and 85.17 percent on the DDR dataset. Hu et al. (2026) address the scale variation and scattered distribution of lesions with a CNN-injected transformer that uses lesion maps twice, once concatenated with the fundus image and once through a dedicated reconstruction branch. Sekar and Raja (2026) fuse an EfficientNet-B3 branch with ViT-B/16 through an attention-augmented feature fusion module, reporting a quadratic weighted kappa of 0.947.

Two observations follow. Attention-weighted pooling and multi-scale fusion are by now established components rather than novel mechanisms, which is why the architecture used here is presented as an experimental vehicle rather than as a contribution. More consequentially, the evidence offered for such components is almost always confined to the distribution on which they were selected.

2.2 Preprocessing, calibration and reliability

Image enhancement is near-universal in DR pipelines. Border removal, field-of-view cropping, illumination correction and CLAHE (Zuiderveld, 1994) are routinely applied, usually justified by the visible improvement in lesion contrast rather than by an ablation isolating the downstream effect. Verma et al. (2026) combine an adaptive fundus enhancement pipeline with a modified EfficientNet-B0 and address overconfidence through test-time augmentation and temperature scaling, reporting 96

percent accuracy and an expected calibration error of 0.030 on IDRiD. Their emphasis on calibration is well placed, and its limits are instructive: Poyrazer et al. (2026) find that temperature scaling reduces development-set calibration error to 0.014–0.022 yet proves largely ineffective under domain shift, with external calibration error rising to 0.086–0.149. Class imbalance is likewise a standard concern, addressed by minority oversampling, class-weighted losses and focal variants, again frequently without a controlled comparison on a fixed partition. The ablations reported below examine both practices under matched conditions.

2.3 Cross-dataset evaluation and ranking stability

Cross-dataset evaluation remains uncommon in DR grading despite being routine in other medical imaging domains, and where it is reported the degradation is substantial. Durmaz Engin et al. (2026) find that models performing strongly under matched conditions degrade once the population changes, with calibration affected as well as discrimination. Pazo et al. (2026) isolate the acquisition device specifically: an EfficientNet-B4 trained on EyePACS and APTOS is applied without retraining first to Messidor-2, captured on a Topcon desktop camera, and then to a portable non-mydriatic device. Kumar et al. (2026) attack the gap architecturally, anchoring crops on a frozen EyePACS teacher's activation map and distilling ordinal CORAL thresholds, and report 0.81 to 0.89 QWK across three unseen datasets.

These studies establish that *absolute* performance falls under transfer. The question examined here is different: whether the *relative ordering* of architectures is preserved. If it is not, the consequence is not merely that deployment figures will fall short of benchmark figures, but that the comparison used to select the architecture may not hold in the setting the model is built for.

Two recent results indicate that this concern is not hypothetical, although neither is presented as such by its authors. Poyrazer et al. (2026) compare three frozen encoders under an identical protocol, developing on APTOS and validating externally on MESSIDOR-2. In distribution the three are indistinguishable, with binary area under the receiver operating characteristic curve (AUC) between 0.980 and 0.985; externally they diverge sharply, to 0.915, 0.745 and 0.697, and the retina-specific encoder falls below the general-purpose ImageNet baseline ($\Delta\text{AUC} = -0.051$, $p = 0.016$). They conclude that development-set discrimination alone is insufficient for encoder selection. Kumar et al. (2026) report, within their own ablation, that the variant achieving the highest in-domain agreement on DDR (0.893, against 0.889 and 0.886) attains the lowest agreement on both external sets (0.693 on MESSIDOR-2 and 0.780 on DeepDRiD), and read this as evidence for their particular crop design rather than as a general property of in-distribution selection.

Neither study treats ordering stability as a hypothesis to be tested. The present work does so directly, varying a single architectural component under otherwise identical conditions and estimating the change in ranking as an interaction with an interval, rather than inferring it from two tables.

3. MATERIALS AND METHODS

3.1 Datasets

Two public datasets are used, both graded on the same five-point international clinical DR severity scale, which makes their labels directly comparable.

APTOS 2019 (Asia Pacific Tele-Ophthalmology Society, 2019) provides 3,662 macula-centred fundus photographs from Indian screening clinics, captured across multiple sites and camera models. It is used for all training and for the primary evaluation.

IDRiD (Porwal et al., 2018) provides 516 graded images acquired in India on a single camera model under conditions distinct from APTOS. It is used exclusively for zero-shot external evaluation and is never seen during training or model selection.

Table 1 and Figure 1 summarise their composition. The two differ markedly in severity distribution: severe and proliferative cases comprise 13.3 percent of APTOS but 30.0 percent of IDRiD, a 2.3-fold difference, so transfer is tested under both appearance shift and prior shift.

Natural Resources for Human Health
eISSN: 2583-1194

Table 1. Dataset composition.

Grade	APTOS 2019 (n)	APTOS 2019 (%)	IDRiD (n)	IDRiD (%)
0 No DR	1805	49.3	168	32.6
1 Mild	370	10.1	25	4.8
2 Moderate	999	27.3	168	32.6
3 Severe	193	5.3	93	18.0
4 Proliferative	295	8.1	62	12.0
Total	3662	100.0	516	100.0
Severe + proliferative	488	13.3	155	30.0

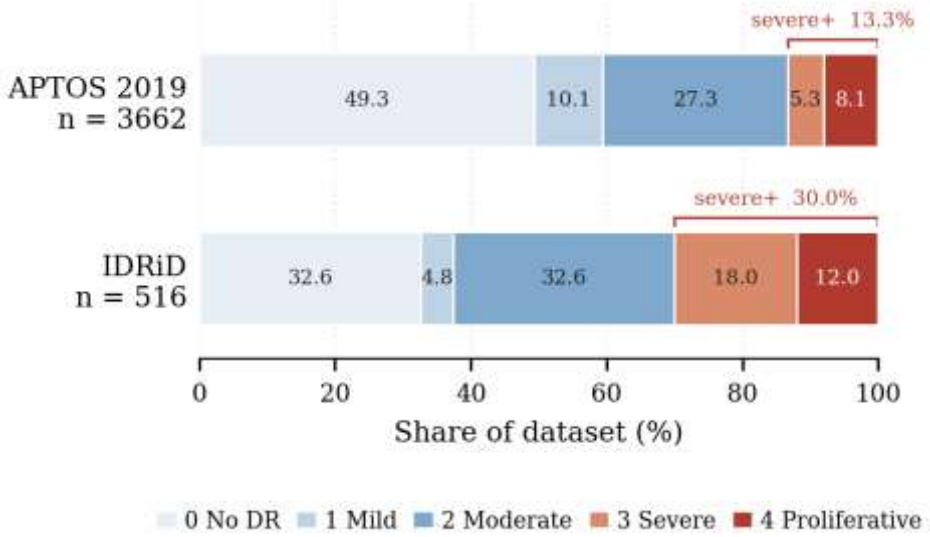


Figure 1. Grade composition of the two datasets. Severe and proliferative cases together account for 13.3 percent of APTOS but 30.0 percent of IDRiD, so the transfer experiment tests a change in class prior as well as a change in imaging conditions.

3.2 Data Partitioning

APTOS is divided once into training, validation and test partitions of 2,563, 549 and 550 images by stratified sampling, preserving class proportions to within 0.03 percentage points. The assignment is written to disk and reloaded by every experiment rather than regenerated from a random seed, so that all comparisons use an identical partition and remain reproducible independently of library version. The test partition is used only for final evaluation.

3.3 Preprocessing and Augmentation

A six-stage preprocessing pipeline was implemented, comprising border removal, retinal field-of-view extraction, illumination correction, CLAHE (Zuiderveld, 1994), adaptive gamma correction and lesion enhancement, each stage building on the previous one. This permits an ablation in which a model is trained on images from different points in the pipeline, with all other factors held constant.

All images are resized to 512 × 512. Training augmentation comprises random horizontal and vertical flips and rotation over the full circle, which are label-preserving for fundus images since they have no canonical orientation. Minority-class samples in

the oversampling condition additionally receive random resized cropping and colour jitter, so that duplicated samples differ from one another (Table 2, Figure 2).

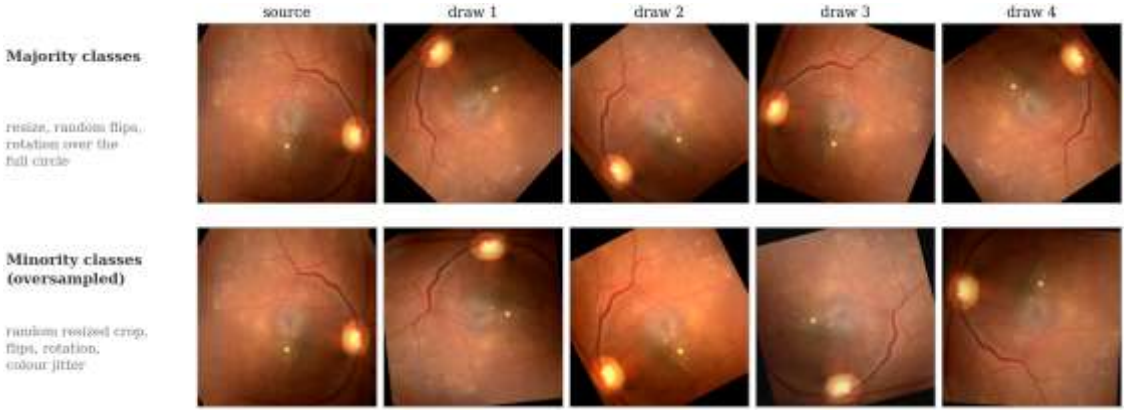


Figure 2. Training augmentation applied to one fundus image, four draws per path. Rotation over the full circle introduces dark corners that do not occur in a real fundus photograph, which is the cost of treating the label as rotation-invariant; the same transform is applied to every variant, so it cannot account for differences between them. In the minority path the crop and colour jitter make duplicated rows differ from their source, which is what distinguishes oversampling from reweighting the loss.

Class rebalancing. Under the oversampling condition, minority-class rows are physically duplicated in the training index by a factor of five for the two most severe grades, following Table 2. Because CaretDataset re-applies the augmentation pipeline on every access, each duplicate receives an independent draw of crop, rotation and colour jitter, so the copies are not pixel-identical and the procedure is not equivalent to simply reweighting the loss. The duplication is applied to the training partition only: validation and test retain the original class proportions, so the reported metrics are measured against the natural distribution (Figure 3).

Table 2. Training set composition before and after class rebalancing.

Grade	Before (n)	Before (%)	Factor	After (n)	After (%)	Val	Test
0 No DR	1263	49.3	1×	1263	32.1	271	271
1 Mild	259	10.1	1×	259	6.6	55	56
2 Moderate	699	27.3	1×	699	17.8	150	150
3 Severe	135	5.3	5×	675	17.2	29	29
4 Proliferative	207	8.1	5×	1035	26.3	44	44
Total	2563	100.0	—	3931	100.0	549	550

Majority-to-minority ratio 9.36:1 before, 4.88:1 after; training set 2,563 to 3,931 images. Validation and test partitions are unaltered.

Natural Resources for Human Health
eISSN: 2583-1194

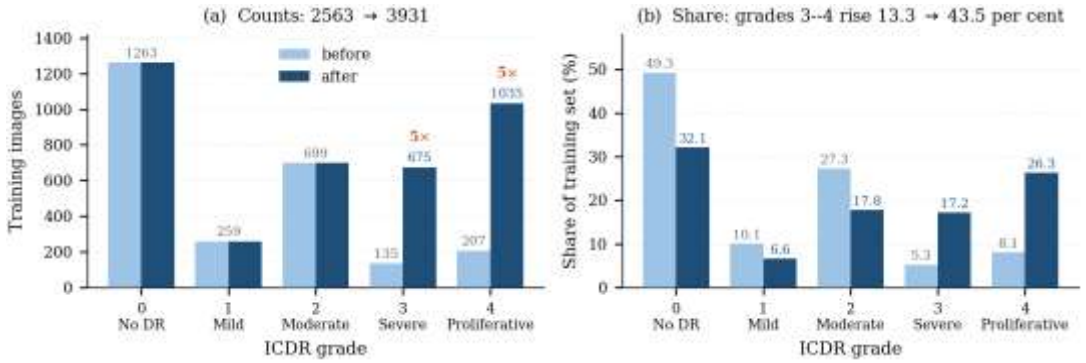


Figure 3. Training-set composition before and after minority oversampling. (a) Counts: the two most severe grades are duplicated fivefold, taking the training partition from 2,563 to 3,931 rows. (b) Shares: the majority grade falls from 49.3 to 32.1 percent while grades 3 and 4 together rise from 13.3 to 43.5 percent, reducing the majority-to-minority ratio from 9.36:1 to 4.88:1. Validation and test partitions are unaltered, so reported metrics are measured against the natural distribution.

3.4 Proposed Architecture

CARETLite couples an EfficientNet-B0 encoder (Tan and Le, 2019), pretrained on ImageNet (Deng et al., 2009), with a lightweight aggregation head. The encoder is fine-tuned end to end; the head replaces the standard global average pooling classifier.

Figure 4 shows the arrangement. Two head components are introduced, each addressing a property of the grading task.

Multi-scale aggregation. DR grades depend on lesions spanning a wide range of sizes, from microaneurysms of a few pixels to diffuse neovascularisation. A classifier reading only the final encoder stage observes a single scale, after successive downsampling has attenuated the smallest structures. The head therefore concatenates pooled descriptors from the final three encoder stages, of 40, 112 and 1280 channels respectively.

Attention pooling. Global average pooling assigns equal weight to every spatial location. Where the diagnostic evidence occupies a small fraction of the field, as when a handful of microaneurysms distinguishes grade one from grade zero, that evidence is attenuated by the surrounding background. Attention pooling instead learns a spatial weighting, following the formulation used for weighted aggregation over instances in multiple-instance learning (Ilse et al., 2018). Given a feature map $\mathbf{X}$ with C channels and spatial dimensions $H \times W$, a two-layer convolutional scoring function produces logits s_{ij}, normalised over spatial positions:

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{p,q} \exp(s_{pq})} \qquad (1)$$

and the pooled descriptor is the weighted sum given in Eq. (2), with the weights of Eq. (1):

$$\mathbf{z}_c = \sum_{i,j} \alpha_{ij}\, \mathbf{X}_{cij} \qquad (2)$$

Variant definitions. The four configurations compared throughout are defined once here and used consistently:

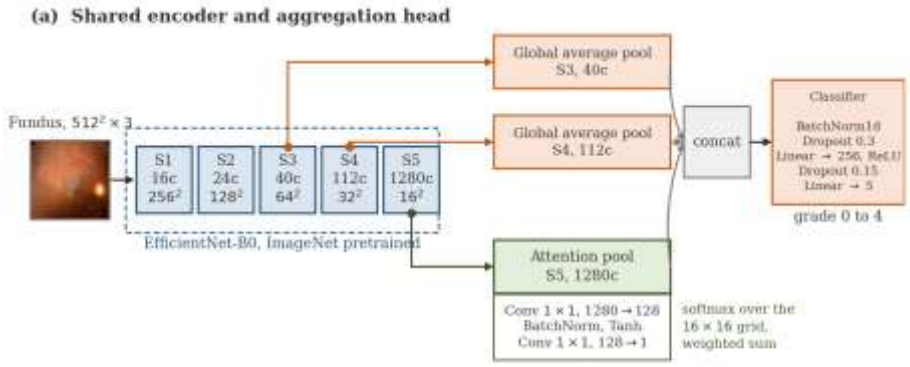
CARETLite-GAP: encoder + global average pooling (GAP) of the final stage. This is the conventional transfer-learning classifier and serves as the control.

CARETLite-MS: encoder + global average pooling of the final three stages, concatenated. This is the multi-scale (MS) aggregation condition.

CARETLite-Attn: encoder + attention (Attn) pooling of the final stage.

CARETLite-MS+Attn: encoder + multi-scale aggregation with attention pooling applied to the deepest stage.

Natural Resources for Human Health
eISSN: 2583-1194



(b) The four compared variants, differing only in the aggregation stage

Variant	S3, S4 taps	S5 pooling	Descriptor	Head par.	Total par.
CARETLite-GAP	–	average	1280-d	0.33 M	4.34 M
CARETLite-MS	yes	average	1432-d	0.37 M	4.38 M
CARETLite-Attn	–	attention	1280-d	0.50 M	4.50 M
CARETLite-MS+Attn	yes	attention	1432-d	0.54 M	4.54 M

Multi-scale aggregation concatenates the pooled S3, S4 and S5 descriptors (40 + 112 + 1280 = 1432); without it only S5 is used (1280). Attention pooling replaces spatial averaging on S5 and adds 0.164 M parameters. Relative to the GAP control the two components together add 0.20 M, 4.5 per cent of the model.

Figure 4. CARETLite. (a) The encoder is a pretrained EfficientNet-B0, shared unchanged across all four variants; channel counts and spatial resolutions are given for a 512 × 512 input, taken as the last feature map before each downsample. The example input is an APTOS training image graded 2, shown at the resolution the network receives. The MS+Attn configuration is drawn: stages S3 and S4 are pooled by spatial averaging and S5 by learned attention, whose scoring function is expanded beneath it, and the concatenated 1432-dimensional descriptor feeds the classifier. (b) The four compared variants. They differ only in whether the S3 and S4 taps are present and in how S5 is pooled; everything else, including the encoder, the classifier and the training protocol, is held fixed. Multi-scale (MS) aggregation and attention (Attn) pooling together add 0.20 million parameters over the global average pooling (GAP) control, 4.5 percent of the model.

CARETLite denotes the family; individual results are always reported against a named variant. No single variant is designated as the proposed model, because the central finding of this paper is that the ranking depends on which dataset the models are evaluated against.

On novelty. The mechanisms drawn on here are established: attention-weighted aggregation over spatial positions (Ilse et al., 2018) and multi-scale feature fusion both have substantial prior literature, and the encoder is an unmodified published backbone (Tan and Le, 2019). The contribution lies in the task-specific composition of these components within a lightweight grading architecture, and in their controlled evaluation, each switched independently, with all other factors held fixed, and each assessed both within and across datasets. The architecture is a means of isolating the aggregation stage, not a claim to a novel mechanism.

Both components are independently switchable, giving four variants: global average pooling alone (GAP, the baseline), multi-scale only (MS), attention only (Attn), and both combined (Figure 5). Each is trained under identical conditions, so the aggregation head is the only deliberately varied architectural factor. Two parameter counts are worth distinguishing, since they are easily confused. The complete aggregation head ranges from 0.33 million parameters in the GAP control to 0.54 million in the combined variant, 7.6 and 11.8 percent of their respective models. What the two components *add* over that control is smaller: total parameter counts range from 4.34 to 4.54 million, so multi-scale aggregation and attention pooling together contribute 0.20 million parameters, 4.5 percent of the model.

Natural Resources for Human Health
eISSN: 2583-1194

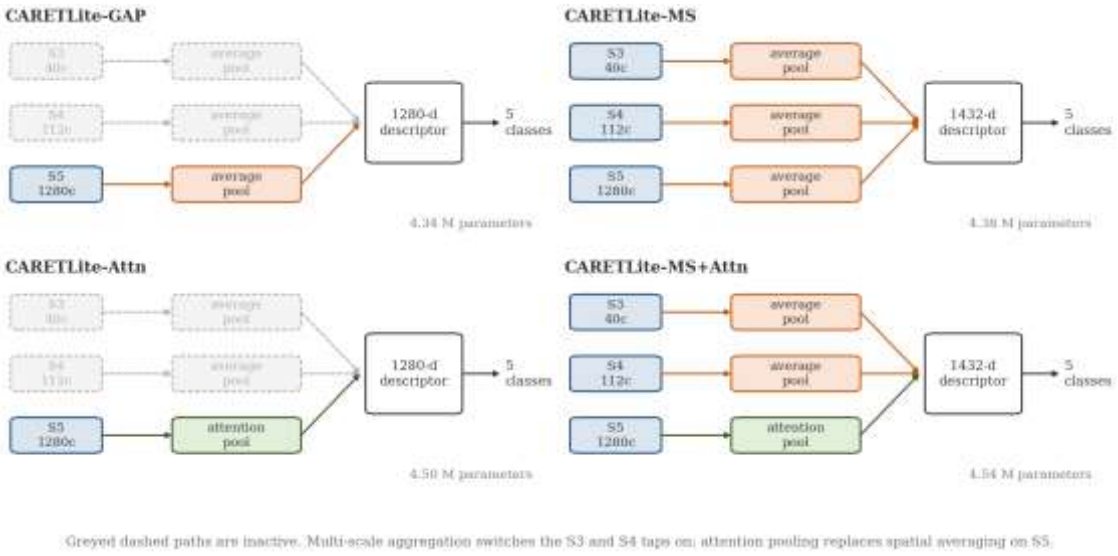


Figure 5. The four compared variants. Each panel shows the same three encoder taps; greyed dashed paths are inactive in that configuration. Multi-scale aggregation switches the S3 and S4 taps on, raising the descriptor from 1280 to 1432 dimensions; attention pooling replaces spatial averaging on S5 and is never applied to the earlier taps. The control sits at top left and the combined variant at bottom right, so the factorial structure reads along the diagonal.

3.5 Training Protocol

All models are trained with Adam at a learning rate of 1×10^{-4} with weight decay 1×10^{-4}, batch size 32, and a maximum of 50 epochs. The learning rate is halved after three epochs without validation improvement, and training stops after seven such epochs. The checkpoint with the lowest validation loss is retained. Cross-entropy loss is used, weighted by inverse class frequency in the class-weighted condition.

Baselines include an EfficientNet-B1 (6.52 million parameters) under the same protocol, and three architectures trained from scratch without pretraining, of 2.97 to 4.85 million parameters, which establish the contribution of pretraining at this data scale.

Parameter counting. All parameter counts in this paper refer to the complete trained model: the convolutional trunk together with the five-class head. The original 1000-class ImageNet classifier is removed in every case, including the EfficientNet-B1 baseline, so the counts are directly comparable. This is why the figures given here are lower than the standard published sizes of these backbones, which include that classifier.

Seeding and repetition. Every configuration reported in Section 4.5 was trained five times, under seeds 42 to 46. The seed is set before the model and the dataloaders are constructed, so it governs head initialisation, shuffling order and the augmentation draws together. The data partition is unaffected: it is read from a frozen file rather than regenerated, so a seed sweep measures training variance alone and not a change of test set. Library versions, hardware and the twenty tracked configuration values are recorded in the output of every run.

Training used Apple Metal (MPS), which does not offer the bitwise determinism available through the deterministic CUDA settings. An unplanned repetition of one configuration nonetheless reproduced its training trajectory exactly, with identical training and validation loss to four decimal places across eleven epochs, and the same early-stopping epoch. The pipeline therefore appears reproducible from a fixed seed in practice, although this is not guaranteed by the platform.

3.6 Evaluation

Choice of primary metric. Quadratic weighted kappa is adopted as the primary metric, and the choice is consequential rather than conventional. Three properties of the grading task motivate it.

First, the grades are *ordinal*. Accuracy, macro-F1 and the standard multi-class scores treat every misclassification identically, so predicting grade 0 for a proliferative case is penalised exactly as heavily as predicting grade 3 for it. That is at odds with the clinical meaning of the scale, where an adjacent-grade disagreement changes a follow-up interval and a four-grade disagreement

changes whether a patient is referred at all. Quadratic weighting penalises a disagreement of d grades in proportion to d^2, so the two are separated by a factor of sixteen. So, let O be the $K \times K$ confusion matrix of counts over the $K = 5$ grades, with O_{ij} the number of images of true grade i predicted as grade j, and let $N = \sum_{ij} O_{ij}$. Quadratic weights and the matrix expected under independent marginals are

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}, \qquad E_{ij} = \frac{1}{N}\left(\sum_q O_{iq}\right)\left(\sum_p O_{pj}\right) \tag{3}$$

and the quadratic weighted kappa is

$$\kappa_w = 1 - \frac{\sum_{ij} w_{ij} O_{ij}}{\sum_{ij} w_{ij} E_{ij}} \tag{4}$$

The numerator is the observed penalty and the denominator is the penalty a predictor reproducing the same marginals by chance would incur, so κ_w is 1 for perfect agreement, 0 for chance-level agreement and negative for agreement worse than chance. The $(i-j)^2$ term is what makes the metric ordinal: an adjacent-grade disagreement carries weight 1/16 of a four-grade one.

Second, the class distribution is heavily skewed, and kappa corrects for agreement expected by chance. The consequence is easily stated: on the test partition used here, a degenerate classifier that assigns grade 0 to every image attains 49.3 percent accuracy but a quadratic weighted kappa of exactly zero (Table 3). Unlike accuracy, kappa adjusts the observed agreement for the agreement expected by chance under the observed marginal distributions. It is not, however, free of the influence of prevalence, and that caveat is relevant here because prevalence differs between the two datasets. This property matters again in Section 4.5, where the class prior differs substantially between the two datasets.

Third, QWK is the metric under which the APTOS 2019 challenge was scored, so reporting it keeps the values commensurable with the body of work built on that dataset.

Table 3. Degenerate predictors on the APTOS test partition.

Predictor	Accuracy	Macro-F1	QWK
Always grade 0 (majority)	0.4927	0.1320	0.0000
Always grade 2 (median)	0.2727	0.0857	0.0000
Random, matching prior	0.3236	0.1930	0.0210

A classifier reproducing the majority class reaches almost half the accuracy of a trained model while carrying no information, which quadratic weighted kappa reflects and accuracy does not.

The distinction is not merely theoretical: the two metrics order the leading models differently in these experiments. Ranked by accuracy the order is EfficientNet-B1 (0.8436), CARETLite-GAP (0.8400), CARETLite-MS (0.8291); ranked by QWK it is CARETLite-MS (0.9070), EfficientNet-B1 (0.9055), CARETLite-GAP (0.8977). The reversal is explained by the distribution of error magnitudes, since CARETLite-MS produces 19 errors of two grades or more against 23 and 25 for the other two. A model can therefore be less often exactly right while being wrong by less, and which of those a metric rewards is a decision the evaluator makes rather than a property of the data.

Metrics reported alongside. Accuracy and macro-averaged F1 are reported throughout because they are widely used and their omission would hinder comparison, and per-class F1 because aggregate scores conceal minority-class behaviour. Unweighted kappa is reported alongside QWK to separate plain agreement from agreement that accounts for how far apart the disagreements are. AUC is not reported because the standard multi-class extensions of it do not use the ordering of the five grades, and so are less suited to the question this paper asks. Mean absolute or squared error over grade indices would respect ordinality but treats the scale as interval-valued and applies no correction for chance agreement.

Natural Resources for Human Health
eISSN: 2583-1194

Interval estimation. Confidence intervals for a single model are obtained by non-parametric bootstrap over the test set (Efron and Tibshirani, 1994). For each of B = 10,000 replicates, n indices are drawn with replacement from the n test cases, QWK is recomputed on the resampled pairs of true and predicted grades, and the reported interval is the 2.5th to 97.5th percentile of the resulting distribution. Percentile intervals are used rather than bias-corrected and accelerated intervals; given the sample sizes involved the distinction is not expected to be material, but the choice is stated so that the figures can be reproduced exactly.

Model comparison. Comparisons between two models use a paired bootstrap. The same B resampling index sets are applied to both models, so each replicate scores both on an identical resample and the difference $\Delta_b = \mathrm{QWK}_b^{(A)} - \mathrm{QWK}_b^{(B)}$ isolates the difference between models from the variance introduced by which cases were drawn. Pairing is possible because every model is evaluated on the same frozen partition, which is verified programmatically before any comparison is made. The reported interval is the 2.5th to 97.5th percentile of $\{\Delta_b\}_{b=1}^{B}$, and the two-sided p-value is $2\min\left(\hat{F}(0),\, 1-\hat{F}(0)\right)$, where $\hat{F}$ is the empirical distribution of Δ_b; that is, twice the smaller of the proportion of replicates in which the difference is non-positive and the proportion in which it is non-negative. A difference is described as significant when the interval excludes zero.

Measuring the change in ranking. Part of this study concerns whether the ordering of two models differs between the two datasets. Observing that model A beats model B on IDRiD, and separately that B beats A on APTOS, does not by itself establish that the gap between them changed; each is a comparison within one dataset. The quantity of interest is the difference of the two differences,

$$D = (\mathrm{QWK}_A - \mathrm{QWK}_B)_{\mathrm{IDRiD}} - (\mathrm{QWK}_A - \mathrm{QWK}_B)_{\mathrm{APTOS}} \qquad (5)$$

which is positive when A gains ground on B under transfer.

D is computed within each seed, from that seed's four scores, and then summarised across seeds by its mean and standard deviation. This is the form reported in Table 8, and it is the form that answers the question, because the variation of interest is between training runs.

For the single-seed analysis of Section 4.6, D was instead estimated by bootstrap with B = 10,000 replicates, resampling the APTOS test partition and IDRiD independently while sharing the drawn index between the two models within each dataset. That procedure is correct for the uncertainty it addresses, which cases were sampled, and its interval excluded zero. It is reported here because it nonetheless failed: the variance that mattered was between training runs, which no resampling of a fixed test set can observe. The contrast between the two estimates of the same quantity is the subject of Section 4.6.

Multiplicity. The seed-level analysis of Section 4.5 computes D for all ten model pairs and reports every one, with no correction applied, because no pair approaches significance and no conclusion rests on any of them; the point of the table is the absence of an effect rather than the presence of one. Where single-seed p-values are quoted, in Section 4.6, and for the preprocessing and imbalance ablations, they are raw bootstrap values over the test partition, and they should be read subject to the limitation established in Section 4.5: an interval of that kind bounds sampling of the test set and not reproducibility across training runs.

4. RESULTS AND DISCUSSION

4.1 Model Comparison Within APTOS

Table 4 reports seed-42 test performance on the 550-image APTOS partition, the single-run view from which this study began. Read on its own it invites a ranking: CARETLite-MS leads at QWK 0.9070 with 4.38 million parameters, ahead of an EfficientNet-B1 baseline at 0.9055 with 6.52 million. Section 4.5 shows that ordering to be a property of the seed rather than of the architectures, and the five pretrained entries should be read as a single cluster. The table is retained because the scratch-trained models in its lower half do separate from that cluster, by a margin examined below.

Natural Resources for Human Health
eISSN: 2583-1194

Table 4. Model comparison on the APTOS test partition.

Model	Par.	Pre.	QWK	95% CI	Acc.
CARETLite-MS	4.38	yes	**0.9070**	[0.881, 0.930]	0.829
EfficientNet-B1	6.52	yes	0.9055	[0.880, 0.929]	0.844
CARETLite-GAP	4.34	yes	0.8977	[0.870, 0.923]	0.840
CARETLite-MS+Attn	4.54	yes	0.8957	[0.865, 0.922]	0.833
CARETLite-Attn	4.50	yes	0.8890	[0.857, 0.917]	0.815
CustomCNN-W	4.85	no	0.8478	[0.814, 0.879]	0.744
CustomCNN-v2	3.38	no	0.8262	[0.788, 0.861]	0.767
CustomCNN	4.85	no	0.8174	[0.779, 0.852]	0.771

Par. = parameters in millions, counted as trunk plus five-class head with the ImageNet classifier removed; Pre. = ImageNet pretraining.

Figure 6 plots this relationship. Pretraining is the one factor examined in this paper whose effect clearly exceeds training noise. Every pretrained model exceeds every scratch-trained model, and the gap from the weakest pretrained run to the strongest scratch-trained one is 0.041 QWK, roughly four times the 0.010 median seed-to-seed standard deviation of Table 7. Three distinct scratch architectures spanning 2.97 to 4.85 million parameters were trained without any reaching 0.85; Figure 7 shows the corresponding convergence behaviour.

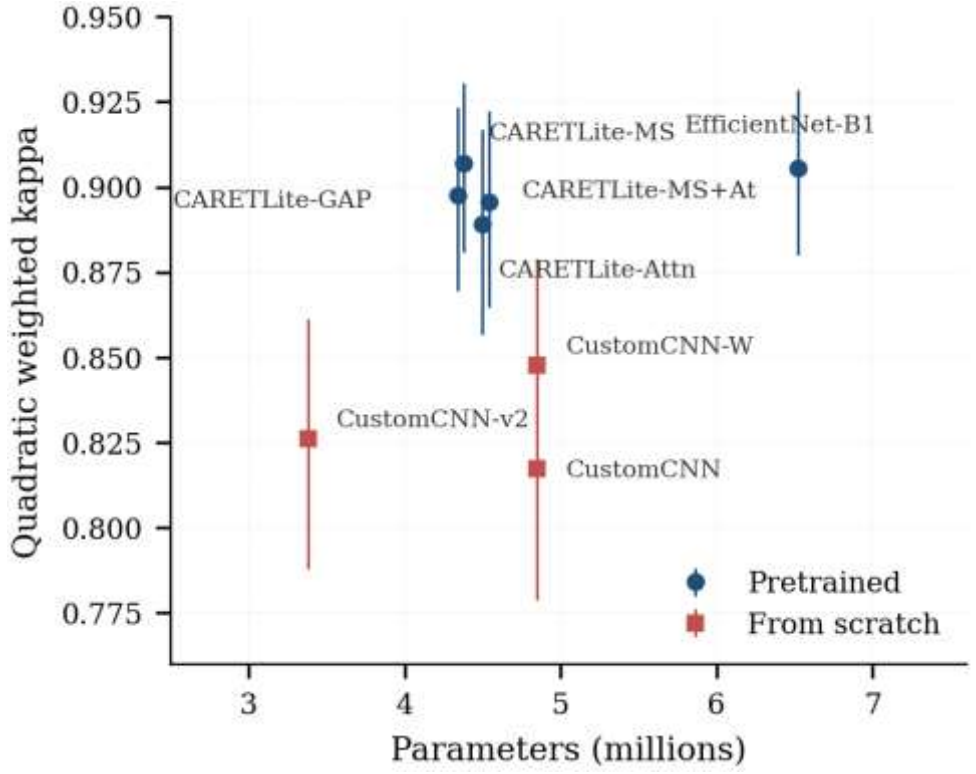


Figure 6. Quadratic weighted kappa against parameter count on the APTOS test partition. Error bars give 95 percent bootstrap intervals. Pretrained models cluster above scratch-trained models irrespective of size.

Natural Resources for Human Health
eISSN: 2583-1194

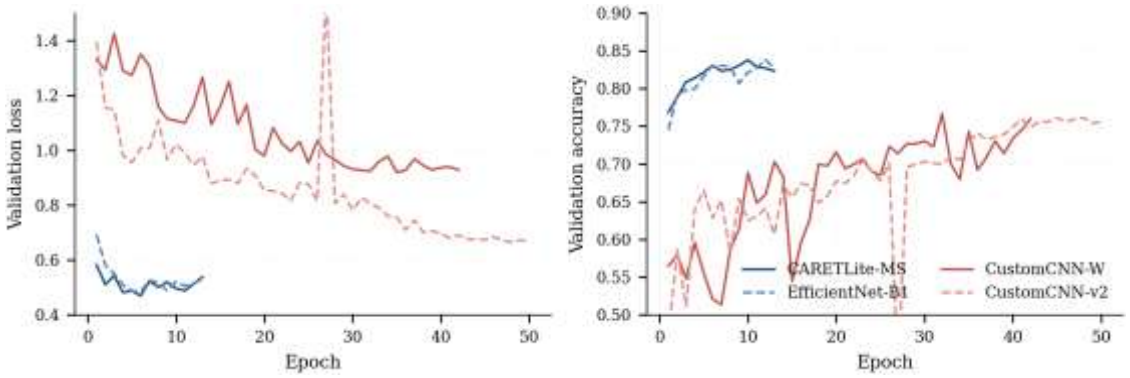


Figure 7. Validation loss and accuracy per epoch for four representative runs. The pretrained models converge earlier, whereas the scratch-trained models require substantially more epochs and reach lower validation performance. Each curve is a single training run.

Two qualifications apply. The comparison is not controlled: pretraining, architecture family and initialisation differ together, so the effect cannot be attributed to pretraining alone. The scratch-trained arm was also run under a single seed, so while the margin is large relative to the seed variance measured in the pretrained arm, it has not itself been replicated. The defensible statement is that the separation between the two groups is several times larger than any architectural difference resolved in this study.

4.2 Head Component Ablation

Table 5 isolates the two head components on seed 42, the configuration in which the experiments were originally run. Within APTOS, multi-scale aggregation improves QWK by 0.0092 over the control and attention pooling reduces it by 0.0086; neither is significant at $n = 550$.

These single-seed differences are reported for completeness and should not be interpreted. Both are smaller than the 0.010 median seed-to-seed standard deviation established in Section 4.5, and across five seeds the mean difference between any two of the four variants is under 0.008 QWK with a standard deviation two to four times larger. The head components examined here have no effect on grading performance that this experiment can resolve.

Table 5. Head component ablation, APTOS test partition, seed 42 only.

Variant	MS	Attn	QWK	Δ	*p*
CARETLite-GAP	—	—	0.8977	—	—
CARETLite-MS	yes	—	0.9070	+0.0092	0.42
CARETLite-Attn	—	yes	0.8890	−0.0086	0.47
CARETLite-MS+Attn	yes	yes	0.8957	−0.0020	0.88

Single seed. All differences are smaller than the seed-to-seed standard deviation of Table 7 and are not interpretable; see Section 4.5.

4.3 Preprocessing Ablation

Table 6 holds model and protocol fixed and varies only the stage of the preprocessing pipeline from which training images are drawn. Unprocessed images give the best result. The full pipeline is indistinguishable from border removal alone, and CLAHE significantly reduces performance relative to raw images ($\Delta = -0.0381$, $p = 0.0016$).

Table 6. Preprocessing ablation, EfficientNet-B1.

Preprocessing stage	QWK	Acc.	Δ	*p*
Raw images	0.9055	0.844	—	—
Border removal	0.8933	0.813	−0.0122	0.35
Full pipeline	0.8935	0.855	−0.0120	0.42
+ CLAHE	0.8678	0.826	−0.0381	**0.0016**

Differences and p-values are relative to raw images.

The result runs counter to common practice. One possible explanation is that CLAHE changes local contrast in a way that makes the model more sensitive to imaging noise and small artefacts, which in a fundus image can resemble the microaneurysms and haemorrhages on which the grading decision depends. A second is that a pretrained encoder already handles some contrast and illumination variation in its early layers, leaving little for the hand-engineered step to add. Neither explanation is tested here; the experiment establishes the effect, not its cause. A further possibility is that the effect depends on the clip limit and tile size chosen, which were held fixed at their default values throughout and not tuned.

4.4 Class Imbalance Handling

Minority oversampling, whose effect on the training composition is given in Table 2, and class-weighted loss were compared with all else fixed. Neither differs significantly for either backbone: $\Delta = -0.030$ ($p = 0.065$) for the scratch-trained CNN and $\Delta = +0.006$ ($p = 0.58$) for EfficientNet-B1. The two strategies do, however, produce different error structures. Class-weighted loss raises severe-class recall from 41 to 62 percent while lowering macro precision from 0.737 to 0.651, an effect the aggregate metrics conceal.

4.5 Seed Variance Against Architectural Difference

Table 7 reports every architecture across five training seeds, on both datasets. The comparison the table is built to support is not between models but between two quantities: how far apart the architecture means are, and how far apart repeated runs of one architecture are.

Table 7. Quadratic weighted kappa across five training seeds.

Dataset	Model	Mean ± SD	Range	Ranks
APTOS	EfficientNet-B1	0.900 ± 0.005	0.893–0.906	1–3
APTOS	CARETLite-GAP	0.895 ± 0.011	0.882–0.907	1–5
APTOS	CARETLite-Attn	0.894 ± 0.010	0.886–0.910	1–5
APTOS	CARETLite-MS+Attn	0.892 ± 0.008	0.878–0.899	2–5
APTOS	CARETLite-MS	0.888 ± 0.011	0.882–0.907	1–5
IDRiD	CARETLite-GAP	0.718 ± 0.028	0.685–0.753	1–4
IDRiD	CARETLite-MS+Attn	0.715 ± 0.019	0.694–0.736	1–5
IDRiD	CARETLite-MS	0.710 ± 0.010	0.701–0.726	1–5
IDRiD	CARETLite-Attn	0.710 ± 0.039	0.657–0.750	1–5
IDRiD	EfficientNet-B1	0.705 ± 0.027	0.663–0.732	2–5

Five seeds per model. "Ranks" gives the best and worst position the model attained across the five seeds, 1 being highest QWK.

Three things follow, and they are visible directly in Figure 8.

Natural Resources for Human Health
eISSN: 2583-1194

The ranking is not stable. Four of the five architectures achieve the highest QWK on at least one seed, on *each* dataset. Six of the ten model-dataset combinations span the full range from first to last. On APTOS the five architecture means occupy a band of 0.012 QWK, and a single architecture moves by up to 0.025 across seeds.

On the external dataset the noise exceeds the signal. The spread between architecture means on IDRiD is 0.013 QWK against a median seed-to-seed standard deviation of 0.027. The variation introduced by re-running the same architecture is twice the largest difference between any two architectures. CARETLite-Attn alone spans 0.657 to 0.750, a range wider than the entire between-model spread.

No pair shows a stable interaction. Table 8 reports the difference-of-differences of Eq. (5) for all ten model pairs, computed within each seed and then summarised across seeds. Every pair changes sign between seeds. The largest mean interaction is +0.018 ± 0.036; the largest t statistic on four degrees of freedom is 2.00 ($p \approx 0.12$). None approaches significance.

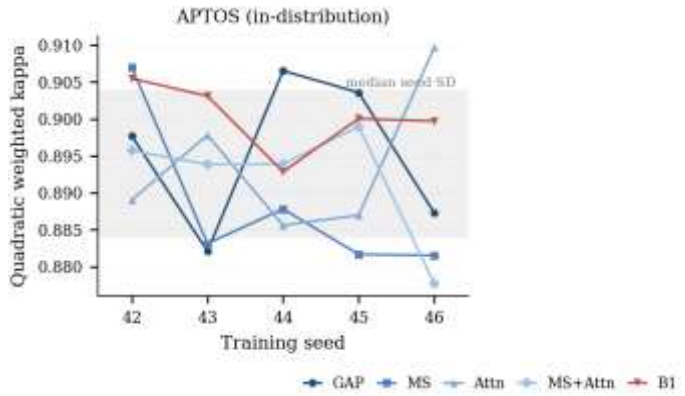


Figure 8. Quadratic weighted kappa against training seed, one line per architecture. The shaded band gives the median seed-to-seed standard deviation of a single architecture, centred on the grand mean, for scale. The lines cross repeatedly: four of five architectures rank first on at least one seed on each dataset. On IDRiD the band is wider than the entire spread between architecture means, so which model appears best in any one column is largely a property of the seed.

Table 8. Architecture-by-dataset interaction across five seeds.

A	B	D mean ± SD	Range	+/5
MS+Attn	B1	+0.018 ± 0.036	−0.020 to +0.074	3
GAP	B1	+0.018 ± 0.031	−0.020 to +0.051	3
MS	B1	+0.017 ± 0.019	−0.011 to +0.042	4
Attn	B1	+0.011 ± 0.053	−0.031 to +0.095	2
GAP	Attn	+0.006 ± 0.036	−0.051 to +0.047	4
MS	Attn	+0.006 ± 0.039	−0.052 to +0.050	3
GAP	MS	+0.000 ± 0.030	−0.038 to +0.031	3
GAP	MS+Attn	−0.001 ± 0.044	−0.038 to +0.071	2
MS	MS+Attn	−0.001 ± 0.033	−0.031 to +0.043	2
Attn	MS+Attn	−0.007 ± 0.041	−0.043 to +0.051	2

D is the change in the A−B QWK gap from APTOS to IDRiD, from Eq. (5), computed within each seed. "+/5" counts seeds with $D > 0$. Every pair changes sign across seeds.

4.6 A Single-Seed Result and Its Refutation

The experiment reported above was prompted by a result obtained before any repetition. On seed 42, CARETLite-Attn ranked last of the four aggregation variants within APTOS (QWK 0.889 against 0.907 for CARETLite-MS) and first on IDRiD (0.742 against 0.707). Testing that as an architecture-by-dataset interaction under the difference-of-differences bootstrap of Eq. (5) gave $D = +0.052$, 95% CI [+0.012, +0.096], $p = 0.012$. The interaction held under six metrics, including accuracy and mean absolute error, neither of which is chance-corrected, which appeared to rule out the class-prior difference between the datasets as an explanation.

Across five seeds that same interaction has mean −0.006 and standard deviation 0.039, the full reversal occurs in one seed of five, and the seed-42 value of +0.052 sits 1.5 standard deviations above a distribution centred on zero. The per-seed values are +0.052, +0.009, −0.025, −0.050 and −0.016: a spread symmetric about nothing.

This is recorded rather than removed because the failure mode is instructive. The original analysis was not casual. It used a pre-specified comparison, a resampling scheme that treated both datasets as random, a three-tier multiplicity structure, and six metrics as a robustness check. None of that helped, because every one of those safeguards addressed uncertainty arising from *which images* were sampled, and the quantity that actually dominated was uncertainty arising from *which training run* was examined. A bootstrap over a test set cannot see that source of variance, however carefully it is constructed.

4.7 Error Structure Under Transfer

Figure 9 plots the difference between predicted and true counts per grade on IDRiD, and identifies the source of the accuracy loss. The pattern is the same for every model: the two sight-threatening grades and the healthy grade are all under-assigned, while the two middle grades absorb the displaced mass. EfficientNet-B1 assigns the proliferative grade to 3 of 62 true cases; the attention variant assigns 22.

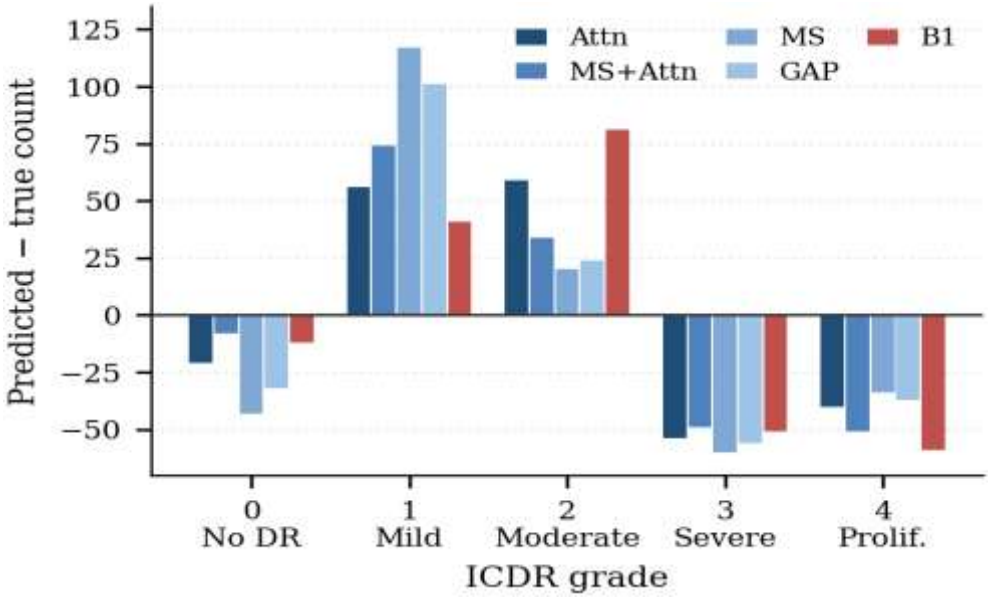


Figure 9. Predicted minus true count per grade on IDRiD. Every model under-assigns severe and proliferative disease and over-assigns the mild and moderate grades, so the degradation under transfer is a systematic compression toward the middle of the scale rather than diffuse error.

The mild grade deserves separate comment, because it fails in a way the aggregate figures conceal. Figure 10 gives per-class F1 within APTOS beside the same quantity on IDRiD. Grade 1 falls from 0.46–0.60 to 0.14–0.18 across every model, with precision near 0.10: the models predict mild disease three to six times more often than it occurs. Two concurrent studies report the same failure. Poyrazer et al. (2026) describe a collapse on mild disease under external evaluation, with two of three encoders reaching F1 exactly zero and the third 0.153, and Kumar et al. (2026) find grade 1 the hardest class on every external set they test. That three independent evaluations converge on the same class suggests the boundary between no retinopathy and mild retinopathy, which rests on the presence of a small number of microaneurysms, is the part of the ordinal scale least robust to a

change of imaging conditions. EfficientNet-B1 additionally reaches F1 exactly zero on proliferative disease under transfer, having assigned that grade to three images in total.

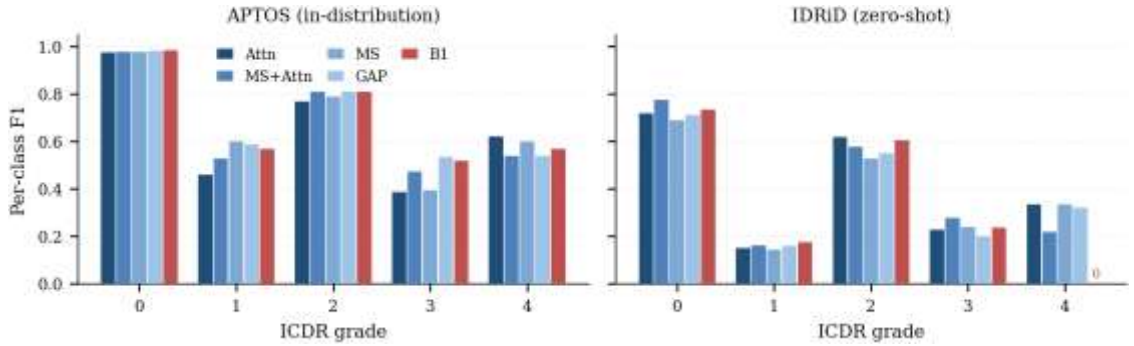


Figure 10. Per-class F1 within APTOS (left) and under zero-shot transfer to IDRiD (right). The mild grade collapses on every model, and EfficientNet-B1 reaches zero on proliferative disease. Aggregate metrics do not reveal either failure.

The observed grade compression is consistent with the combined effects of prevalence shift and other dataset-specific distributional differences. The two datasets differ in camera, acquisition protocol and image characteristics as well as in prevalence, and the present experiment does not separate these factors; attributing the compression specifically to prevalence shift would require prior correction or recalibration against the target distribution, which is left to future work. Severe and proliferative cases constitute 13.3 percent of APTOS but 30.0 percent of IDRiD. This difference in class prevalence is one plausible contributor to the observed compression, but its effect cannot be separated from other dataset-specific differences in the present experiment. Whatever its cause, the consequence for a screening application is direct and adverse: sight-threatening disease is systematically under-assigned in a setting where its prevalence is higher than in the training data. The effect is invisible to the aggregate metrics and entirely invisible to single-dataset evaluation, and it is not small: the largest baseline identified 3 of 62 proliferative cases. Any deployment of a grading model in a population differing from its training data should therefore be accompanied by an explicit check of the predicted severity distribution against the expected prevalence.

4.8 Reported Results in Prior Studies

Table 9 lists quadratic weighted kappa values reported for APTOS 2019 in a selection of recent studies. It is included with an explicit caveat, because presenting such a table sits in tension with this paper's own argument. The studies use different train and test partitions, and several report performance on the challenge leaderboard rather than on a locally held-out set, so the values are not measured on common data and no ranking should be inferred from them. More importantly, the results of Section 4.5 indicate that a single-dataset figure is a weak predictor of behaviour under domain shift, which applies to the entries in this table as much as to the present work. The table is therefore offered to situate the magnitude of the numbers involved, not to establish a comparison.

Table 9. Reported results on APTOS 2019 in selected prior studies.

Study	Method	Backbone size	QWK	Ensemble	External val.	CI / sig. test
Chilukoti et al. (2024)	EfficientNet-B3, transfer + ensemble	~12M × n	0.967	yes	no	no
Shakibania et al. (2024)	Dual-branch EfficientNet-B0 + ResNet50	~29M	0.930	yes	no	no
Dharrao et al. (2025)	EfficientNet-B0 + multi-scale attention	~5M	0.923	no	no	yes

Study	Method	Backbone size	QWK	Ensemble	External val.	CI / sig. test
This work	**EfficientNet-B0 + multi-scale head**	**4.38M**	**0.907**	**no**	**yes**	**yes**
Dharrao et al. (2025)	DenseNet121 baseline	~8M	0.908	no	no	yes
Dharrao et al. (2025)	ResNet50 baseline	~25M	0.901	no	no	yes
Marrero Ortiz (2026)	EfficientNet-B3 + explainable AI	~12M	0.879	no	no	no
Al-Smadi et al. (2021)	DenseNet-169 + Inception-V3 + Xception	~60M	0.824	yes	no	no

Partitions and evaluation protocols differ between studies; values are not measured on common test data and should be read as context rather than as a controlled comparison. Backbone sizes are approximate, taken from the standard architecture definitions where not stated by the authors. Entries marked "yes" under CI / sig. test report standard deviation over repeated runs rather than confidence intervals on the test statistic.

Two observations are worth drawing from it. The highest reported values are obtained by ensembles, which multiply inference cost by the number of constituent models. Among single-model methods the values cluster between 0.88 and 0.93, and the best variant reported here falls within that band while using the smallest parameter budget in the table. None of the listed studies reports evaluation on an independent dataset.

The contribution of this work is therefore not a higher in-distribution score. It is the observation that such scores, reported in isolation, do not determine which architecture should be preferred — a claim the present experiments support directly, and one that applies to the values in Table 9 as much as to those in Table 4.

4.9 Error Severity

Because QWK penalises errors by squared distance, the distribution of error magnitudes is informative. On APTOS, CARETLite-MS places 82.9 percent of predictions exactly and 96.6 percent within one grade, with no error exceeding three grades. EfficientNet-B1 attains higher exact accuracy, 85.5 percent, but produces more large errors, with seven predictions off by three grades against three for CARETLite-MS. This accounts for CARETLite-MS achieving the higher QWK despite the lower accuracy, and reflects a different error profile, whose desirability depends on the referral thresholds of the screening programme in question: an adjacent-grade error alters a follow-up interval, whereas an error of three grades may return a sight-threatening case to routine screening.

4.10 Discussion

The central result is a null one with a practical consequence. Across twenty-five training runs, the differences between five architectures are not larger than the differences between repeated runs of any one of them. On the external dataset they are half as large. Whichever architecture appears best in a given experiment is, at this data scale, substantially determined by the random seed.

That has an immediate implication for how this literature reports results. The comparisons that motivate architectural choices in DR grading are typically of the size measured here, one to two hundredths of a QWK, and are typically reported from a single run. Section 4.5 indicates that such comparisons carry little information. This is not a claim that the components examined are useless in general; it is a claim that an experiment of this size cannot tell whether they help, and that reporting one run as though it could is misleading.

The same concern applies to concurrent work. Kumar et al. (2026) compares three architectures under identical settings and report that the variant leading in-domain on DDR trails on both external sets, attributing this to a fixed-crop overfitting mode; that ablation is run under a single seed. Poyrazer et al. (2026) find three frozen encoders indistinguishable in distribution and sharply divergent externally, with the retina-specific encoder falling below the ImageNet baseline; their external comparison likewise rests on one training of each head. Both results may well be real. The present work cannot adjudicate them, but it does show that an effect of that magnitude, obtained from a single run in this domain, is within the range that seed variation alone produces. The appropriate response is replication rather than reinterpretation.

A well-built significance test did not protect against this. The refuted analysis of Section 4.6 used a pre-specified comparison, an interaction rather than two separate rankings, independent resampling of the two evaluation sets, an explicit multiplicity structure and a six-metric robustness check. Each of those addresses uncertainty about which images were drawn. None addresses uncertainty about which training run was examined, and in this regime the second dominates. Bootstrap intervals over a test partition should not be read as bounding the reproducibility of a model comparison.

What survives replication is the error structure. The degradation under transfer is systematic and identical in shape across every model and every seed: predictions compress toward the middle of the ordinal scale, the sight-threatening grades are under-assigned, and mild disease collapses. These are properties of the transfer itself rather than of any architecture, which is why they are stable where the rankings are not. The mild-grade collapse in particular is corroborated independently by both Poyrazer et al. (2026) and Kumar et al. (2026), and it points at a specific weakness: the boundary between no retinopathy and mild retinopathy rests on a small number of microaneurysms, and that evidence does not survive a change of imaging conditions.

Pretraining is the one factor that clears the noise. The separation between pretrained and scratch-trained models is roughly four times the observed seed variance, against architectural differences at or below one times. Where an effect is large relative to training noise it can be resolved by a small experiment; where it is not, no amount of careful analysis of a single run will substitute for repetition.

4.11 Limitations

The seed sweep covers five seeds and five architectures on one training dataset. Five is a small sample for estimating a standard deviation, and the standard deviations reported in Table 7 are themselves uncertain; they should be read as establishing the order of magnitude of training variance rather than its precise value. The conclusion drawn from them, that architectural differences of one to two hundredths of a QWK cannot be resolved by a single run, is robust to that imprecision, since it would hold for any plausible value in the range observed.

The pre-processing and class-imbalance ablations were run under a single seed and inherit the same limitation. The CLAHE result ($\Delta = -0.038$) is roughly four times the APTOS seed standard deviation and so may well survive repetition, but that has not been tested here and it is reported as provisional. The imbalance comparison was null under one seed and remains untested.

The scratch-trained baselines were likewise run once. The pretraining separation is large enough that seed variance is unlikely to account for it, but the claim rests on seed variance measured in the pretrained arm.

External evaluation employs 516 images from a single additional dataset and camera. Whether the rank instability reported here is specific to this pair of datasets, to this architecture family, or to this data scale is not determined by these experiments. The most useful extension would be to repeat the measurement at a larger training-set size, where architectural differences may grow relative to training noise.

5. CONCLUSION

Five architectures, differing in feature aggregation and in backbone size, were each trained five times under identical conditions on APTOS 2019 and evaluated both in distribution and, with weights frozen, on IDRiD. The differences between architectures are not larger than the differences between repeated runs of the same architecture: 0.012 QWK of spread against a median seed-to-seed standard deviation of 0.010 on APTOS, and 0.013 against 0.027 on IDRiD. Four of the five architectures rank first on at least one seed on each dataset, and most span the full range from first to last. No model pair shows a stable architecture-by-dataset interaction.

A single-seed version of this experiment produced an apparently significant architectural reversal between the two datasets ($D = +0.052$, $p = 0.012$, holding under six metrics). It did not replicate: across five seeds the same interaction has mean -0.006

and standard deviation 0.039. That sequence is reported in full because the original analysis was carefully constructed and still described noise, every safeguard in it addressed sampling of the test set, while the dominant variance came from training.

Two findings survive replication. Degradation under transfer is systematic rather than diffuse, with predictions compressing toward the middle of the ordinal scale and mild disease collapsing from an F1 of 0.46–0.60 to 0.14–0.18 on every model and every seed. And ImageNet pretraining separates from scratch training by roughly four times the observed seed variance, making it the only factor examined whose effect this experiment can resolve.

The recommendation is narrow and practical. Architectural comparisons in lightweight DR grading at this data scale should be accompanied by repeated training, or the differences they report should not be used to choose between architectures. Where repetition is not feasible, the seed variance of the setting should at least be measured once, so that a reader can judge whether a reported difference exceeds it.

AUTHOR CONTRIBUTIONS

Himansu Sekhar Rout: Conceptualisation, Methodology, Software, Investigation, Writing—Original Draft. Saravana Kumar S: Supervision, Validation, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study used two publicly available, fully anonymised retinal fundus image datasets (APTOS 2019 and IDRiD). No new data were collected from human participants and no identifiable personal information was accessed, so institutional ethical approval was not required.

DATA AVAILABILITY STATEMENT

Both datasets used in this study are publicly available: APTOS 2019 from the Kaggle competition repository (https://www.kaggle.com/c/aptos2019-blindness-detection) and IDRiD from https://doi.org/10.3390/data3030025. The data partitions, trained model configurations and per-image predictions supporting the reported results are retained by the authors and are available on reasonable request to permit independent verification.

REFERENCES

[1] Al-Smadi, M., Hammad, M., Baker, Q.B., Al-Zboon, S.A., 2021. A transfer learning with deep neural network approach for diabetic retinopathy classification. International Journal of Electrical and Computer Engineering 11(4), 3492–3501. https://doi.org/10.11591/ijece.v11i4.pp3492-3501

[2] Asia Pacific Tele-Ophthalmology Society, 2019. APTOS 2019 blindness detection. Kaggle. https://www.kaggle.com/c/aptos2019-blindness-detection

[3] Chilukoti, S.V., Shan, L., Tida, V.S., Maida, A.S., Hei, X., 2024. A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric. BMC Medical Informatics and Decision Making 24, 37. https://doi.org/10.1186/s12911-024-02446-x

[4] Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70(4), 213–220. https://doi.org/10.1037/h0026256

[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[6] 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

[7] Durmaz Engin, C., Selver, M.A., Köksaldı, S., Kayabaşı, M., 2026. Generalizability of deep learning models for referable diabetic retinopathy detection: A cross-population study. Retina. https://doi.org/10.1097/IAE.0000000000004924

[8] Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780429246593

[9] 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. https://doi.org/10.1001/jama.2016.17216

[10] Hu, Y., Li, H., Xu, Q., Lin, Y., Shi, J., Shen, J., Xu, Y., Dou, X., 2026. A CNN-injected transformer network with lesion reconstruction for multi-view diabetic retinopathy grading. Medical Image Analysis 113, 104178. https://doi.org/10.1016/j.media.2026.104178

[11] Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning, in: Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80, pp. 2127–2136. https://proceedings.mlr.press/v80/ilse18a.html

[12] Kumar, P., Singh, S.K., Rawat, U., 2026. Cross-dataset diabetic retinopathy grading with a lesion-anchored multi-view transformer and ordinal-distance knowledge distillation. ICT Express. https://doi.org/10.1016/j.icte.2026.06.010

[13] Marrero Ortiz, J.A., 2026. Integrated decision support system for diabetic retinopathy grading using explainable AI and uncertainty quantification. Graduate Project Article, Graduate School, Polytechnic University of Puerto Rico. https://hdl.handle.net/20.500.12475/3375

[14] Pazo, E.E., Liu, X., Liu, S., Zhang, Q., Zhang, M., Zhang, Z., Moutari, S., Usama, M., Ren, X., 2026. Validation of the Eyerobo FC portable fundus camera for diabetic retinopathy screening using public datasets and deep learning. Ophthalmology and Therapy 15, 1439–1461. https://doi.org/10.1007/s40123-026-01353-w

[15] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F., 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD): A database for diabetic retinopathy screening research. Data 3(3), 25. https://doi.org/10.3390/data3030025

[16] Poyrazer, M., Yağcı, H., Erten, R., 2026. How well do frozen foundation models transfer? A calibration-focused benchmark for diabetic retinopathy grading. Frontiers in Medicine 13, 1815982. https://doi.org/10.3389/fmed.2026.1815982

[17] Sekar, P., Raja, K.S., 2026. AI driven hybrid CNN transformer model for early detection and severity assessment of diabetic retinopathy. Scientific Reports. https://doi.org/10.1038/s41598-026-50608-w

[18] Shakibania, H., Raoufi, S., Pourafkham, B., Khotanlou, H., Mansoorizadeh, M., 2024. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. Biomedical Signal Processing and Control 93, 106168. https://doi.org/10.1016/j.bspc.2024.106168

[19] Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, pp. 6105–6114.

[20] Verma, P., Elango, S., Singh, K., 2026. DR-TrustNet: Enhancing diabetic retinopathy detection using reliable efficient networks and uncertainty quantification. Image and Vision Computing 168, 105921. https://doi.org/10.1016/j.imavis.2026.105921

[21] Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., et al., 2003. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 110(9), 1677–1682. https://doi.org/10.1016/S0161-6420(03)00475-5

[22] Yi, S., Yang, C., Ma, L., 2026. L-MedViT: A hierarchical hybrid network for fine-grained diabetic retinopathy grading. Image and Vision Computing 170, 105970. https://doi.org/10.1016/j.imavis.2026.105970

[23] Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization, in: Heckbert, P.S. (Ed.), Graphics Gems IV. Academic Press, San Diego, pp. 474–485. https://doi.org/10.1016/B978-0-12-336156-1.50061-6