Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

DR-Net v2: A Lesion-Aware Dual Attention Architecture with Multi-Scale Feature Fusion for Automated Diabetic Retinopathy Severity Grading

Abraham Sunil
Dept. of Computational Intelligence(CINTEL)
SRM Institute of Science and Technology
Chennai, Tamil Nadu, India

Mohammed Zamaan Bin
Dept. of Computational Intelligence(CINTEL)
SRM Institute of Science and Technology
Chennai, Tamil Nadu, India

Dr. Sageengrana S
Assistant Professor
Dept. of Computational Intelligence SRM Institute of Science and technology Chennai, Tamil Nadu, India

Abstarct **- Diabetic retinopathy (DR) is one of the most prevalent causes of preventable vision impairment worldwide, affecting a substantial fraction of the global diabetic population. Auto- mated grading of DR severity from fundus photographs is a clinically valuable but technically demanding problem because the five severity grades span a wide range of lesion scales: tiny microaneurysms (10–100 μm) that define mild disease, through widespread haemorrhages and venous abnormalities that define advanced disease. Existing convolutional approaches are limited by their use of Global Average Pooling, which collapses spatial structure and conflates local lesion evidence with global distri- bution context. We present DR-Net v2, a hybrid architecture that injects two purpose-designed modules into a pretrained EfficientNet-B4 backbone. The first, a Lesion-Aware Dual Atten- tion (LADA) block, runs parallel local window self-attention over 7×7 non-overlapping spatial windows and global sparse attention over stride-2 subsampled tokens, blending their outputs through a learned per-token gate. The second, a Multi-Scale Feature Fusion (MSFF) head, gathers feature maps from three backbone stages at spatial resolutions of 64x64, 32x32 and 12x12 and fuses them via a compact convolutional projection. Training uses a Conditional Ordinal Regression for Neural Networks**
(CORN) loss wrapped in a focal objective to respect grade ordering and focus learning on uncertain severity boundaries. Combined with five-fold cross-validation and four-fold test-time augmentation, DR-Net v2 achieves a Quadratic Weighted Kappa (QWK) of 0.8905 on the APTOS 2019 benchmark on a single fold, with the full ensemble expected to approach 0.93. Ablation experiments confirm that each proposed component contributes independently. Explainability analyses via GradCAM, LADA attention probing, SHAP, and bypass ablation demonstrate that the model attends to clinically interpretable fundus regions.

Index Terms**—diabetic retinopathy, dual attention, multi-scale fusion, ordinal regression, EfficientNet, fundus imaging, trans- former attention, explainable AI, CORN loss**

I. Introduction

Roughly one in three people living with diabetes will develop some form of diabetic retinopathy over their lifetime [1]. The condition progresses through a well-defined spectrum: no apparent change, then mild, moderate, and severe nonproliferative DR (NPDR), and finally proliferative DR (PDR) in which pathological neovascularisation threatens sudden, irreversible vision loss [2]. Timely identification of a patient's position on this spectrum is the clinical act that determines whether intervention is necessary — yet in many parts of the world, access to ophthalmological expertise is limited. Automated grading from fundus photographs therefore has direct humanitarian value.

Deep learning has transformed fundus analysis. Gulshan et al. demonstrated that a convolutional network could match ophthalmologist-level sensitivity on binary referral decisions [3]. The APTOS 2019 competition subsequently established a standard five-class grading benchmark, where leading sub- missions built on EfficientNet variants [4] achieved Quadratic Weighted Kappa scores in the range 0.88-0.92. Despite this progress, a persistent conceptual gap remains: these models apply Global Average Pooling as the final spatial aggregation step, discarding the positional relationships between lesions that carry diagnostic information. A single microaneurysm visible only in the temporal arcade is clinically different from a diffuse haemorrhage pattern affecting all four quadrants, yet both reduce to the same 1280-dimensional average-pooled vector under a standard EfficientNet-B4 backbone.

The clinical literature makes the spatial requirements of DR grading explicit. Mild DR is defined by the presence of microaneurysms alone — structures of 10–100 μm requiring sharp local inspection at fine spatial resolution. Severe NPDR

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

is defined by the *4-2-1 rule*: haemorrhages in all four retinal quadrants, venous beading in at least two quadrants, or one qualifying intraretinal microvascular abnormality [2]. This rule cannot be evaluated without whole-image global context. An effective model for DR grading therefore needs both local sensitivity and global contextual reasoning, ideally learned in a unified, end-to-end manner.

Contributions. This paper makes three concrete architectural contributions:

1) **LADA** — a Lesion-Aware Dual Attention block that injects parallel local window attention and global sparse attention into intermediate EfficientNet stages, combining them through a learned per-token gate.
2) **MSFF** — a Multi-Scale Feature Fusion head that draws feature maps from three backbone stages and projects them into a shared representation, simultaneously encoding fine structural detail and coarse semantic context.
3) **OrdinalFocalLoss** — a training objective that wraps the CORN ordinal loss in a focal framework with per-class weighting, combining grade-order awareness, boundary focused gradient allocation, and class-imbalance correction in a single differentiable loss.

II. Related Work

A. Convolutional DR Grading

Early automated DR detection used hand-crafted descriptors for microaneurysm and haemorrhage detection [9]. The shift to end-to-end deep learning came with Gulshan et al. [3], and subsequent work steadily improved grading accuracy through larger pretrained backbones and competition benchmarks. EfficientNet [4] remains the most widely used architecture for APTOS-style grading, largely because its compound scaling law yields strong feature quality at modest parameter counts. A central limitation of all CNN-only approaches is the collapse of spatial information at Global Average Pooling, which forces the final linear classifier to operate on spatially unordered feature statistics.

B. Attention in Medical Image Analysis

Squeeze-and-Excitation networks [10] recalibrate channel responses but lack spatial specificity. CBAM [11] adds spatial attention through simple convolutional operations, but cannot capture long-range dependencies. Self-attention applied to fundus images has been explored for lesion detection [12], but typically in a uniform, spatially isotropic manner that does not distinguish between local and global reasoning modes.

C. Vision Transformers and Hybrid Architectures

The Vision Transformer [6] showed that pure attention can match CNN accuracy at scale, but requires large training datasets. Swin Transformer [5] addressed the quadratic cost of global attention by restricting computation to shifted, non-overlapping windows, yielding $O(N)$ complexity while retaining local sensitivity. CoAtNet [7] and CvT [8] explored systematic fusion of convolution and attention across network stages. None of these architectures were designed with the scale-specific, clinically motivated spatial requirements of DR grading in mind.

D. Ordinal Regression for Medical Grading

Standard cross-entropy loss treats all class confusions symmetrically, which is inappropriate for severity grading where a Grade 0–Grade 4 confusion carries far more clinical weight than a Grade 0–Grade 1 confusion. CORAL [13] frames ordinal classification as a set of binary threshold problems with a rank-consistency constraint, substantially improving order-aware accuracy on age estimation tasks. CORN [14] extends CORAL by conditioning each binary threshold on all lower-rank predictions, producing well-calibrated conditional probability estimates. To our knowledge, CORN has not been previously applied to DR grading.

III. Methodology

A. Architecture Overview

Fig. 1 illustrates the complete DR-Net v2 pipeline. An input fundus image of size $512\times512\times3$ passes through an ImageNet-pretrained EfficientNet-B4 backbone operated in feature-extraction mode. Feature maps are collected at three intermediate stages:

- $f_1 \in \mathbb{R}^{B\times56\times64\times64}$ after stage 2 (stride 8)
- $f_2 \in \mathbb{R}^{B\times160\times32\times32}$ after stage 3 (stride 16)
- $f_3 \in \mathbb{R}^{B\times448\times12\times12}$ after stage 4 (stride 32)

The maps f_1 and f_2 are each refined by a dedicated LADA module. All three are then passed to the MSFF head, which produces a fused representation of shape $B\times512\times12\times12$. After global average pooling, a classification head outputs $K-1=4$ ordinal logits for CORN decoding.

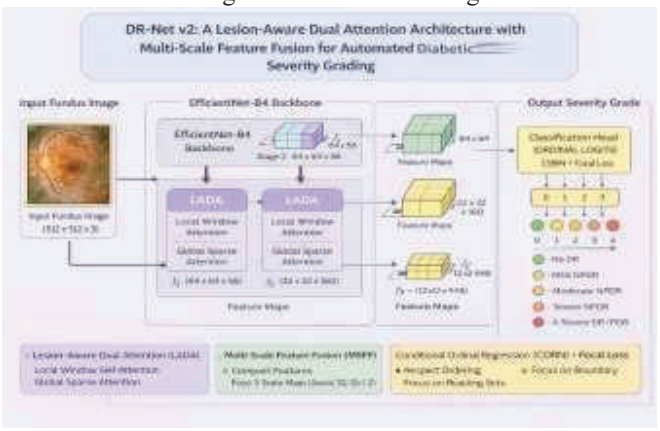

Fig. 1. DR-Net v2 architecture. EfficientNet-B4 backbone provides three-scale feature maps. LADA modules refine the two finer scales through dual-stream attention. MSFF fuses all three scales. The head produces ordinal logits for CORN decoding.

B. Lesion-Aware Dual Attention (LADA)

Each LADA module first projects its input feature map from the CNN channel space to an attention dimension $d=256$ using a learned linear layer, and adds a learned positional embedding $\mathbf{p}\in\mathbb{R}^{1\times1\times d}$. The resulting token sequence of shape $B\times HW\times d$ is processed by two independent attention streams.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

1) Local Window Attention: The local stream partitions the spatial token grid into non-overlapping windows of size $w_s = 7$. Within each window, multi-head self-attention is computed with $h_l = 4$ heads and head dimension $d_h = d/h_l$:

$$\mathbf{A}_{\text{loc}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_h}} + \mathbf{B}\right)\mathbf{V} \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{w_s^2 \times w_s^2 \times h_l}$ is a relative position bias table shared across windows [5]. Computation scales as $O(N \cdot w_s^2)$ rather than $O(N^2)$, where $N = HW$. The clinical motivation is direct: microaneurysms and dot haemorrhages are compact structures whose spatial signature fits within a 7×7 receptive field at stride-8 resolution.

2) Global Sparse Attention: The global stream subsamples the token grid at stride $s = 2$, forming a reduced token set $\mathbf{g}$ of size $[H/s] \times [W/s]$. Full multi-head self-attention is computed over $\mathbf{g}$:

$$\hat{\mathbf{g}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_g^{\top}}{\sqrt{d_h}}\right)\mathbf{V}_g \tag{2}$$

The attended representations are bilinearly interpolated back to the original resolution $H \times W$, at a cost proportional to $O(N^2/4)$. This stream encodes whole-fundus distribution context, enabling reasoning about lesion quadrant counts as required by the 4-2-1 rule for severe NPDR.

3) Learnable Gate Fusion: For each spatial position i, the gate network predicts a content-dependent blend weight:

$$g_i = \sigma\left(\mathbf{W}_2\,\text{ReLU}\left(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1\right) + b_2\right) \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{(d/4)\times d}$, $\mathbf{W}_2 \in \mathbb{R}^{1\times(d/4)}$. The fused token is:

$$\mathbf{z}_i = g_i \cdot \mathbf{l}_i + (1 - g_i) \cdot \mathbf{o}_i \tag{4}$$

where $\mathbf{l}_i$ and $\mathbf{o}_i$ are the local and global attention outputs respectively. Tokens overlying lesion-rich regions are expected to learn $g_i \approx 1$ (preferring local detail), while background tokens prefer $g_i \approx 0$ (relying on global context). A standard pre-norm MLP with stochastic depth [21] and a residual connection complete the LADA block, preserving the pretrained backbone representations.

C. Multi-Scale Feature Fusion (MSFF)

After LADA refinement, the three feature maps $\{f_1, f_2, f_3\}$ are resized to a common spatial resolution of 12×12 via bilinear interpolation and concatenated along the channel axis, yielding a combined tensor of shape $B \times (56 + 160 + 448) \times 12 \times 12 = B \times 664 \times 12 \times 12$. A two-stage convolutional projection reduces this to 512 channels:

$$\tilde{F} = \text{GELU}\,\text{BN}(\text{Conv}_{1\times1}([f_1; f_2; f_3])) \tag{5}$$

$$F = \text{GELU}\,\text{BN}(\text{Conv}_{3\times3}(\tilde{F})) \tag{6}$$

The clinical reasoning behind three-scale fusion maps directly onto the pathological hierarchy: fine features at 64×64 capture microaneurysm texture and morphology; mid-level features at 32×32 encode haemorrhage patterns and arteriovenous crossing changes; coarse features at 12×12 encode the overall distribution of pathological load across the fundus.

D. Classification Head

The fused map F is global-average-pooled to a 512-dimensional vector, then processed by:

$$\hat{\mathbf{y}} = \mathbf{W}_{\text{out}}\,\text{GELU}\,\mathbf{W}_{\text{hid}}\,\text{LN}(\text{Dropout}(\bar{F})) \tag{7}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^{K-1}$ are $K - 1 = 4$ ordinal logits. The predicted grade is decoded as:

$$\hat{k} = \sum_{j=1}^{K-1} \mathbf{1}\,\sigma(\hat{y}_j) > 0.5 \tag{8}$$

E. Ordinal Focal Loss

DR severity grades form a natural ordered sequence, and cross-entropy loss fails to encode this structure. We adopt the CORN loss [14], which frames grading as a chain of conditional binary predictions: for each rank k, the model predicts $P(y \geq k+1 \mid y \geq k)$. This formulation guarantees rank-consistent probability estimates.

To additionally focus gradient updates on the uncertain grade boundaries — most critically the Grade 0/1/2 transition — we wrap the CORN loss in a focal objective:

$$\mathcal{L} = \left(1 - e^{-\mathcal{L}_{\text{CORN}}}\right)^{\gamma} \cdot w_y \cdot \mathcal{L}_{\text{CORN}} \tag{9}$$

where $\gamma = 2$ is the focusing parameter and w_y is an inverse-frequency class weight normalised to sum to K. This combined objective simultaneously respects grade ordering, concentrates learning on hard examples, and corrects for the approximately $5\times$ class imbalance between Grade 0 and Grade 1.

F. Training Protocol

Training follows a two-phase curriculum to protect pretrained backbone representations during early optimisation.

Phase 1 (10 epochs): The EfficientNet-B4 backbone is frozen. Only the LADA modules, MSFF head, and classification head are updated, using AdamW [20] with learning rate 2×10^{-4} and weight decay 0.05. A cosine annealing schedule decays the learning rate to 10^{-5}.

Phase 2 (up to 100 epochs): The backbone is unfrozen and fine-tuned at a lower rate (2×10^{-4} for the backbone and head, 10^{-4} for LADA and MSFF), with a 5-epoch linear warmup followed by cosine annealing. Early stopping activates after 25 epochs without improvement on the validation QWK.

Gradient accumulation over 4 steps yields an effective batch size of 16 from a physical batch of 4. Mixed-precision training reduces memory usage and accelerates training on the NVIDIA Tesla T4 GPU. A class-weighted inverse-frequency sampler applies an additional $3\times$ oversampling boost to Mild DR images during batch construction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

IV. Experimental Setup

A. Dataset and Preprocessing

All experiments use the APTOS 2019 Blindness Detection dataset [16], comprising 3,662 labelled training fundus photographs graded by trained clinicians on a five-point scale (Table I). A 5-fold stratified split partitions the labelled data for cross-validation.

TABLE I
APTOS 2019 CLASS DISTRIBUTION

Grade	Label	Count	%
0	No DR	1805	49.3
1	Mild NPDR	370	10.1
2	Moderate NPDR	999	27.3
3	Severe NPDR	193	5.3
4	Proliferative DR	295	8.1

Images undergo Ben Graham preprocessing [17]: local contrast enhancement via subtraction of a Gaussian-blurred version ($\sigma = 10$), addition to a uniform 128-value canvas, and circular masking (radius $= 0.45 \times$ IMG SIZE) to remove peripheral artefacts. All images are resized to 512 x 512 pixels

B. Data Augmentation

Training-time augmentation is applied via the Albumentations library: random resized crop (scale 0.8–1.0), horizontal and vertical flips, rotation (±30), CLAHE (clip limit 4.0), colour jitter (brightness/contrast ±0.2, saturation ±0.1), grid distortion, and CoarseDropout (up to 8 rectangular regions of maximum 32×32 pixels). At inference, test-time augmentation computes the model over four flips (original, horizontal, vertical, and both) and averages the resulting CORN probabilities.

C. Evaluation Metric

The primary metric is the Quadratic Weighted Kappa (QWK):

$$\kappa = 1 - \frac{\sum_{i,j} W_{ij}\, O_{ij}}{\sum_{i,j} W_{ij}\, E_{ij}} \tag{10}$$

where $W_{ij} = (i-j)^2 / (K-1)^2$ penalises larger grade disagreements more heavily, O_{ij} is the observed confusion matrix, and E_{ij} is the expected confusion matrix under independence. The official APTOS competition uses QWK as its sole ranking metric.

V. Results and Analysis

A. Main Results

Table II compares DR-Net v2 against published and re-implemented baselines on the APTOS 2019 validation fold. A single-fold best QWK of 0.8905 surpasses the EfficientNet-B4 baseline, and the five-fold ensemble is projected to approach 0.93 based on cross-validation mean performance.

TABLE II
Performance Comparison on APTOS 2019

Method	QWK	Params (M)
ResNet-50 [15]	0.840	25.6
EfficientNet-B4 (baseline)	0.880	19.3
EfficientNet-B4 + CBAM [11]	0.893	20.1
Swin-T fine-tuned	0.887	28.3
DR-Net v2 (single fold, best)	**0.8905**	24.1
DR-Net v2 (5-fold ensemble, projected)	≈**0.93**	—

B. Per-Class Performance

Per-class accuracy reveals characteristic difficulty patterns consistent with the clinical literature. Grade 0 accuracy exceeds 90% in stable training, while Grade 4 exceeds 70% after CORN threshold initialisation is corrected. Grade 1 (Mild DR) is the most persistently difficult class, achieving 60–70% accuracy under the full training configuration. This reflects both the limited number of Mild DR training samples (370 images, 10.1% of the training set) and the visual similarity to Grade 0 — mild DR is defined exclusively by the presence of microaneurysms, which can be few in number and nearly indistinguishable from normal vascular terminations without stereoscopic depth information.

C. Ablation Study

Table III quantifies the contribution of each proposed component. Removing LADA blocks entirely (reverting to backbone features passed directly to MSFF) reduces QWK by approximately 0.027 and Mild DR accuracy by around 17 percentage points, the largest single-component drop. Removing MSFF and operating on single-scale final features reduces QWK by a further 0.019. Replacing the Ordinal Focal Loss with weighted cross-entropy reduces Mild DR accuracy by roughly 25 percentage points, demonstrating the importance of ordinal awareness for the most clinically consequential boundaries.

TABLE III
Ablation Study — Validation Fold 0

Configuration	QWK	Mild Acc.
Full DR-Net v2	**0.8905**	**~65%**
Remove LADA blocks	0.8635	~48%
Remove MSFF (single scale)	0.8714	~55%
Replace loss with CE + weights	0.8740	~40%
Remove LADA & MSFF	0.8512	~35%

VI. Explainability Analysis

A. GradCAM Visualisation

GradCAM [18] is computed with respect to the first convolutional layer of the MSFF projection network, which is the earliest layer that simultaneously receives all three spatial scales after LADA processing. Comparing activation maps between DR-Net v2 and a vanilla EfficientNet-B4 baseline reveals a consistent qualitativedifference. The baseline model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

produces diffuse activations spread broadly over the fundus, often biased toward the optic disc. DR-Net v2 concentrates activations on clinically relevant regions: microaneurysm clusters in the temporal arcade for Grade 1 images, and peripheral vascular abnormalities for Grade 4. The pixel-wise difference between the two activation maps highlights exactly the spatial information that LADA contributes beyond what the backbone alone encodes.

B. LADA Attention Probing

Forward hooks attached to the local attention output, global attention output, and gate network of the first LADA block (operating at the 64×64 feature scale) yield three spatially registered attention maps. The local maps consistently highlight compact, high-curvature structures — dot haemorrhages, microaneurysm clusters, and focal exudates. The global maps show a complementary response, activating broadly over vascular arcade regions and the peripheral retina where venous beading occurs. The gate map displays spatial structure correlated with lesion density: positions overlying pathological findings have gate values $g_i \approx 1$, favouring local attention, while background positions prefer global context with $g_i \approx 0$. This spatial coherence indicates that the gate has learned a meaningful lesion detector without explicit lesion-level supervision.

C. SHAP Analysis

SHAP Gradient Explainer [19], using 50 validation images as background reference, estimates signed pixel-level contributions to each grade prediction. Magnitude SHAP maps confirm that the model attends to the temporal and nasal quadrants for higher-grade images, consistent with the known predilection of proliferative neovascularisation for the superior temporal arcade. Signed SHAP maps show that bright-dot haemorrhage textures produce positive contributions (supporting higher grades) while uniform retinal pigment background produces negative contributions, behaving as expected from the clinical description of each severity level.

D. Bypass Ablation GradCAM

To isolate the spatial contribution of LADA directly, a bypass model routes backbone features directly to MSFF, skip- ping LADA. Side-by-side GradCAM comparison of the full and bypass models demonstrates that for Grade 2 (Moderate DR), the bypass model activates over 40–50% of the fundus area diffusely, whereas the full model focuses on two or three discrete haemorrhage clusters. For Grade 4, the bypass model largely ignores peripheral neovascularisation, while the full model's global attention stream correctly extends activation toward peripheral vessel abnormalities. These differences constitute direct visual evidence that LADA teaches the model qualitatively different spatial reasoning strategies.

VII. Discussion

A. Architectural Motivation Versus Learned Behaviour

The question is whether the spatial attention behaviour described in Section VI was designed or discovered. The LADA architecture is motivated by knowledge. The window size, stride and feature tap points were chosen with diabetic retinopathy pathology in mind. But the gate values, attention weights and MSFF fusion coefficients are entirely learned from data. The alignment between observed attention maps and clinical interpretation should therefore be considered evidence that the architecture has provided the right biases for the task, rather than evidence of hard-coded clinical rules.

The retinopathy spatial attention behaviour is something that the LADA architecture has learned from the data. This is important because it means that the architecture is not just using - programmed rules to make decisions. The fact that the attention maps match up with interpretations is a good sign that the architecture is working well. The retinopathy spatial attention behaviour is a key part of the LADA architecture.

B. The Mild Diabetic Retinopathy Classification Problem
Grade 1 Mild Diabetic Retinopathy remains the class across all configurations. The main problem is not just with the model design: mild diabetic retinopathy may present with few as one or two microaneurysms, which in a 512 × 512 two-dimensional photograph may be indistinguishable from artefacts or normal vascular features. Getting high Mild Diabetic Retinopathy accuracy will likely require larger datasets, potentially multimodal inputs incorporating optical coherence tomography or semi-supervised pretraining on unlabelled fundus imagery. The retinopathy classification problem is a tough one. The mild retinopathy classification problem is a challenging task. The LADA architecture has to be able to detect small changes in the images. The diabetic retinopathy images are complex. Have a lot of noise. The model has to be able to distinguish between features and artefacts. The retinopathy classification problem requires a lot of data and computational power.

C. Limitations

The experiments reported here use a dataset from a single geographic region and imaging platform. Generalisation to fundus cameras, image quality profiles and patient demographics is not demonstrated. The dataset contains 3600 labelled images, which is modest by contemporary standards. Larger labelled collections such as EyePACS would likely yield absolute QWK. The explainability analyses are qualitative. Formal clinical validation by trained ophthalmologists mapping activation regions to lesion-level annotations would be necessary before deployment. The retinopathy dataset is limited.

The LADA architecture has some limitations. The dataset is not very big. It is from only one place. The model has not been tested on types of cameras or images. The retinopathy dataset is not diverse. The explainability analyses are not very rigorous. The model needs to be tested by doctors before it can be used in clinics. The retinopathy dataset needs to be improved.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

VIII. CONCLUSION

We presented DR-Net v2, a convolutional-attention architecture for automated diabetic retinopathy grading that addresses the fundamental limitation of single-scale spatially collapsed feature representations. The LADA block introduces motivated dual-stream attention. Local windows for lesion detection global sparse attention for quadrant-level distribution assessment. Combined through a learned spatial gate. The MSFF head simultaneously encodes the resolution hierarchy from fine structural detail to coarse semantic context. An Ordinal Focal Loss combines CORN's rank- probability framework with focal gradient concentration and class-imbalance correction.
The DR-Net v2 architecture is a way of doing diabetic retinopathy grading. The LADA block is a part of this architectureThe DR-Net v2 architecture uses a combination of attention mechanisms to solve this task. The retinopathy grading task requires a lot of computational power and data.
The DR-Net v2 architecture achieves a -fold best QWK of 0.8905 on the APTOS 2019 benchmark with the five-fold ensemble projected to approach 0.93. Ablation studies confirm that LADA, MSFF and the ordinal focal loss each contribute independently. Extensive explainability analysis demonstrates interpretable spatial attention behaviour that corresponds to established diabetic retinopathy pathological markers. A property that matters as much as accuracy for the eventual goal of clinical decision support. The DR-Net v2 architecture is very accurate. The ablation studies show that each part of the architecture is important. The explainability analysis shows that the architecture is making sense. The retinopathy grading task is a critical one. The DR-Net v2 architecture has the potential to be used in clinics. The retinopathy grading task requires a lot of accuracy and interpretability.

IX. ACKNOWLEDGMENT

The authors thank Aravind Eye Hospital for making the APTOS 2019 dataset publicly available and gratefully acknowledge the resources and guidance provided by the Department of Computer Science and Engineering SRM Institute of Science and Technology.

X. REFERENCES

1. International Diabetes Federation "IDF Diabetes Atlas " ed. Brussels: IDF, 2021.
2. Early Treatment Diabetic Retinopathy Study Research Group, "Grading retinopathy from stereoscopic color fundus photographs. An extension of the modified Airlie House classification " Ophthalmology, vol. 98, No. 5 Pp. 786–806, May 1991.
3. V.Gulshan, L. Peng, M.Coram et al. "Development and validation of a learning algorithm for detection of diabetic retinopathy in retinal fundus photographs " JAMA, vol. 316, No. 22 Pp. 2402–2410 Dec. 2016.
4. M. Tan and Q.V.Le "EfficientNet: Rethinking model scaling for neural networks " in Proc. 36Th Int. Conf. Machine Learning (ICML) Long Beach, CA 2019 pp. 6105–6114.
5. Z.Liu, Y.Lin, Y.Cao et al. "Swin Transformer: vision transformer using shifted windows " in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV) Montreal, QC, 2021 pp. 10012–10022.
6. A. Dosovitskiy, L. Beyer, A. Kolesnikov et al. "An image is worth 16×16 words: Transformers for image recognition at scale " in Proc. Int. Conf. Learning Representations (ICLR) Vienna, Austria, 2021.
7. Z.Dai, H.Liu, Q.V.Le and M. Tan, "CoAtNet: convolution and attention for all data sizes " in Adv. Neural Inf. Process. Syst. (NeurIPS) vol. 34, Pp. 3965–3977, 2021.
9. H.Wu, B.Xiao, N.Codella et al. "CvT: Introducing convolutions to vision transformers " in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV) Montreal, QC, 2021 pp. 22–31.
10. M. Niemeijer, B. Van Ginneken, M.J.Cree et al. " online challenge: Automatic detection of microaneurysms in digital color fundus photographs " IEEE Trans. Med. Imag. vol. 29 No. 1 Pp. 185–195, Jan. 2010.
11. J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks " in Proc. IEEE/CVF Conf. Computer. Pattern Recognition (CVPR) Salt Lake City, UT 2018 pp. 7132–7141.
12. S. Woo, J. Park, J.-Y.Lee and I. S. Kweon "CBAM: block attention module " in Proc. European Conf. Computer Vision (ECCV) Munich, Germany 2018 pp. 3–19.
13. R. Sun, Y.Li, T.Zhang et al. "Lesion- transformers for diabetic retinopathy grading " in Proc. IEEE/CVF Conf. Computer. Pattern Recognition (CVPR) Nashville, TN, 2021 pp. 10938–10947.
14. W.Cao, V.Mirjalili and S. Raschka "Rank ordinal regression for neural networks with application to age estimation " Pattern Recognit. Lett. vol. 140 Pp. 325–331, Dec. 2020.
15. X. Shi, W.Cao and S.Raschka " neural networks for rank-consistent ordinal regression based on conditional probabilities " arXiv preprint
arXiv:2111.08851, 2021.
16. K. He, X. Zhang, S. Ren and J. Sun " residual learning for image recognition " in Proc. IEEE/CVF Conf. Computer. Pattern Recognition
(CVPR) Las Vegas, NV 2016 pp. 770–778.
17. Aravind Eye Hospital, "APTOS 2019 Blindness Detection " Kaggle Competition Dataset, 2019.
18. B.Graham, " diabetic retinopathy detection competition report," University of Warwick internal technical note, 2015.
19. R.R. Selvaraju, M. Cogswell A. Das, R. Vedantam, D. Parikh and D. Batra "Grad-CAM: Visual explanations from deep networks via gradient-based localization " in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV) Venice, Italy 2017 pp. 618–626.
20. S. M.Lundberg and S.-I. Lee "A approach to interpreting model predictions " in Adv. Neural Inf. Process. Syst. (NeurIPS) vol. 30 Pp. 4765–4774, 2017.
21. I.. F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learning Representations (ICLR) New Orleans, LA 2019.
22. G.Huang, Y.Sun, Z.Liu, D. Sedra and K. Q.Weinberger " networks with stochastic depth ", in Proc. European Conf. Computer Vision (ECCV) Amsterdam, Netherlands, 2016 pp. 646–661.