

Deployment Aligned Hybrid Attention For Diabetic Retinopathy Screening With Mdrnet

Reji R¹, Gripsy Paul Mannickathan², Suresh H³, Vijilesh Krishnan V K⁴, Biku Abraham⁵, Sujith Kumar P S⁶, Thomas Joseph⁷, Jubilant J Kizhakkethottam⁸, Lekshmi V S⁹, Chinchu M S¹⁰, Divya Saleela¹¹

¹Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

²Assistant Professor, CMR University, Karnataka, India

³Assistant Professor, Thangal Kunju Musaliar Institute of Technology, India

⁴Associate Professor, KMCT Institute of Technology and Management, Kerala, India

⁵Professor, Saintgits College of Engineering, Kerala, India

⁶Professor, Sree Buddha College of Engineering, Kerala, India

⁷Assistant Professor, Carmel College of Engineering and Technology, Kerala, India

⁸Professor, SRM Institute of Science and Technology, Chennai, India

⁹Assistant Professor, Sree Buddha College of Engineering, Kerala, India

¹⁰Assistant Professor, Sree Buddha College of Engineering, Kerala, India

¹¹Assistant Professor, College of Engineering Aranmula, Kerala, India

Abstract

Early identification of diabetic retinopathy (DR) is critical for preventing avoidable vision loss, particularly in primary-care settings where screening infrastructure is constrained by variable image quality, limited hardware resources and workforce pressure. This work introduces MDRNet, a deployment-aligned deep learning framework designed to provide clinically coherent DR grading under these real-world conditions. The approach integrates F-CAN perceptual harmonisation, a lightweight EfficientNet-B0 encoder, multi-scale hybrid attention, and class-distribution regularisation, yielding a model tailored for commodity CPU hardware without reliance on transformer-scale capacity. Evaluation on a patient-level EyePACS split that preserves natural epidemiological imbalance demonstrates robust multiclass performance, achieving 91.3% accuracy, 0.86 macro-F1, 0.90 quadratic weighted kappa, and 0.95 macro-AUC across the five-grade clinical scale. Referral-tier mapping, which collapses severe and proliferative disease into a clinically actionable category, achieved 0.92 sensitivity and 0.85 specificity, indicating suitability as a triage-supportive system. Calibration metrics (Brier score = 0.082; ECE = 0.034) further suggest reliability of probabilistic outputs for threshold-based decision support. The findings demonstrate that prioritising calibration, ordinal stability and inference feasibility can yield competitive performance without specialised accelerators, supporting the potential for MDRNet to contribute meaningfully to scalable DR screening pathways subject to prospective validation.

Keywords : Diabetic retinopathy, retinal disease screening, hybrid attention networks, class-imbalance learning.

1. Introduction

Diabetic retinopathy is a leading cause of avoidable visual impairment and blindness, and its global prevalence continues to rise in line with increasing diagnoses of diabetes mellitus and demographic ageing. Estimates indicate that more than one hundred and forty five million individuals were affected in 2021, with forecasts surpassing two hundred and twenty four million by 2030 [1]. The disease frequently progresses without symptoms, and irreversible pathological changes to retinal structure may occur before clinical attention is sought. As a result, the capacity of healthcare systems to provide timely screening has become integral to preventing sight loss, particularly in regions where ophthalmic services are unevenly distributed. The effectiveness of screening programmes, however, is shaped by limitations in resources, personnel and image quality, all of which complicate the early identification of pathological change.

Clinical assessment typically relies on colour fundus photography and subsequent grading by trained specialists [2]. Although this approach remains the clinical standard, its performance depends on consistent conditions of image acquisition, illumination and focus, which are not always achievable in practice. Subtle pathological indicators such as microaneurysms and small haemorrhages can be concealed by acquisition variability or sensor limitations, and this hinders the identification of the earliest clinically relevant stages of the disease [3] [4]. Diagnostic performance also varies across graders, and reliability is influenced by workload pressure and the availability of second opinions [5]. These constraints illustrate a structural reliance on human expertise that cannot scale indefinitely, particularly where screening demand exceeds capacity.

Machine learning research has presented automated grading systems as a strategy to expand screening capability [6]. Convolutional models and representation learning approaches have demonstrated encouraging results on benchmark datasets, and recent studies report levels of agreement with human graders that approach clinical acceptability under controlled experimental conditions [7]. However, translation into routine practice has remained limited. Many models exhibit sensitivity to domain shift, including differences in camera hardware, imaging environments and population characteristics. Furthermore, substantial reliance on graphical processing hardware restricts deployment to well-resourced settings and introduces dependence on cloud connectivity [8]. The operation of such systems beyond tertiary care centres is therefore constrained, despite the fact that the greatest clinical benefit of automation may lie in precisely those contexts where infrastructure is limited.

Dataset structure presents an additional barrier. Screening populations demonstrate statistically skewed distributions, with mild and moderate presentations substantially more common than severe or proliferative cases. Standard optimisation procedures are not inherently robust to this imbalance and may bias predictions towards the majority categories, resulting in high reported accuracy alongside clinically unacceptable misclassification patterns [9]. In response, research has begun to adopt evaluation strategies that reflect medical risk rather than aggregate correctness, including macro averaged measures and ordinal agreement statistics that penalise errors in a manner that mirrors clinical consequence. These developments indicate a gradual shift in emphasis from headline performance metrics to the practical reliability required for decision support.

Clinical integration also depends on interpretability. Automated output that cannot be interrogated or scrutinised by a practitioner is unlikely to satisfy current standards for accountability or regulatory approval [10]. The ability to demonstrate that decision-making mechanisms correspond to identifiable retinal structures and established pathological pathways forms an increasingly explicit requirement. This has situated explainability as a parallel objective alongside accuracy and operational feasibility. In practice, a system must not only perform well on test data but must do so in a manner that can be examined, questioned and validated by clinicians.

This study is situated within these conditions and is directed towards automated diabetic retinopathy grading that acknowledges the practical realities of screening environments. The objective is to strengthen performance for the earliest stages of disease, to moderate the influence of skewed data distributions, and to maintain feasibility for operation on central processing units typical of primary or community care [11]. Rather than refining performance exclusively for curated datasets or computationally intensive hardware, the work is motivated by alignment with clinical pathways in which screening volume is high, acquisitions are heterogeneous and access to specialist devices or infrastructure is inconsistent.

The contribution of this study is positioned in the development of an automated approach to diabetic retinopathy grading that prioritises sensitivity to early pathological change, robustness under statistical imbalance and compatibility with widely available hardware. The intention is to advance the field in the direction of systems that are better suited to the demands and constraints of population-level screening, with relevance to settings in which present technological expectations cannot be met.

The remainder of the article provides a contextual review of relevant literature; a specification of the methodological framework developed in response to the challenges outlined above and an evaluation on the EyePACS dataset using clinically representative measures. The article concludes with a discussion of findings, limitations and implications for future research and deployment.

2. Literature Review

Automated diabetic retinopathy analysis has developed across several established directions that reflect distinct priorities in the research community. Broadly, the literature can be located within studies that extend convolution based classification, investigations that pursue resilience to variation in acquisition settings, approaches designed to counter skewed distributions of disease severity, contributions focused on interpretability and accountability, and work driven by the practical need for deployment on central processing units in non specialist healthcare environments. These directions provide a useful conceptual structure through which to examine existing progress and outstanding limitations.

Convolution driven approaches were central to the initial establishment of automated grading as a credible research agenda. Models based on Inception, ResNet and comparable architectures reported strong results on curated datasets such as EyePACS and Messidor, demonstrating marked advantages over handcrafted descriptors and shallow classifiers [12] [13] [14]. These contributions remain influential benchmarks. However, their effectiveness has been shown to depend on consistent illumination, controlled focus and uniform data preparation, and they typically exhibit reduced sensitivity to early pathological cues when image quality deviates from these conditions. This indicates a persistent deficit in approaches oriented toward detectability of early subtle lesions under clinically representative variation.

Research concerned with generalisation across acquisition domains acknowledges the variability introduced by differences in camera hardware, imaging protocol and patient demographics. Techniques including domain invariant representation learning, alignment of feature distributions and augmentation targeted at stylistic variation have been proposed to moderate performance degradation when training and deployment contexts diverge [15] [16]. These methods demonstrate improved robustness under selective forms of domain shift. Nonetheless, many assume access to graphical acceleration or controlled optimisation procedures that are unrealistic in community screening contexts, and thus a gap remains in solutions that maintain domain stability while remaining computationally feasible on central processing units.

Class frequency imbalance represents a defining characteristic of screening datasets, where severe and proliferative cases are comparatively rare despite being critical for treatment decisions. Strategies such as class reweighting, modified optimisation dynamics and ordinal formulations have been introduced to reinforce minority categories and reduce systematic underestimation of severity [17] [18]. Although these contributions report numerical improvement, evaluation frameworks have frequently relied on aggregate accuracy or top-1 correctness, which does not convey clinical consequence when models misgrade borderline severity. This highlights the need for imbalance responsive approaches evaluated under metrics that align with clinical decision

thresholds and risk profiles, including macro averaged measures and ordinal agreement statistics.

Interpretability has become increasingly central as automated systems approach clinical consideration. Attribution mechanisms, saliency mapping and lesion level visualisation have been employed to support practitioner scrutiny and to align predicted outcomes with anatomically plausible retinal features [19] [20]. These developments respond to ethical, regulatory and epistemic expectations. However, interpretability is often appended after core model design rather than informing it, and attribution behaviour remains unstable when input quality deteriorates. This suggests that further work is required to deliver interpretability that is resilient to imperfect imagery and capable of contributing to clinical accountability within everyday screening practice.

A more recent literature engages directly with infrastructural constraints, reflecting the fact that primary care providers, mobile screening units and community health programmes frequently rely on central processing units and do not possess consistent access to specialist hardware or cloud inference pathways. These contributions reduce architectural scale, prioritise reduced energy cost or eliminate reliance on graphical acceleration [21], [22]. While this represents a significant step toward practicality, these works typically consider computational feasibility independently from the challenges of early sensitivity, domain variation and imbalance, limiting their suitability for scaled utilisation in real screening pathways.

Taken together, the literature demonstrates important progress in diagnostic performance, domain stability, imbalance mitigation, accountability and resource efficiency. However, these developments remain largely fragmented. The evidence indicates that existing systems either lose sensitivity to early pathological indications when image quality varies, or address imbalance without adopting clinically aligned evaluation criteria, or obtain robustness through computational strategies that presuppose graphical hardware, or incorporate interpretability as a retrospective addition that is not stable under non ideal input conditions. As a consequence, there remains a critical gap in the formulation of automated screening systems that treat early lesion detectability under variable acquisition, imbalance aware learning assessed through clinically relevant metrics, feasibility on central processing units and resilient interpretability as interdependent requirements rather than discrete objectives. The methodological section that follows is articulated directly in response to this gap and outlines a framework constructed to operate under the empirical and infrastructural conditions characteristic of routine diabetic retinopathy screening.

3. Methodology

The proposed methodology is formulated to align algorithmic design with the operational realities of DR screening in primary-care pathways. Rather than optimising for peak performance under laboratory conditions, the framework is constructed around three governing principles: fidelity to the clinical data distribution, architectural efficiency compatible with commodity CPU hardware, and optimisation strategies that respect the ordinal and imbalanced structure of DR severity scales. These principles act as design priors that shape each stage of the pipeline, from data curation and perceptual harmonisation to representation learning and diagnostic evaluation. The resulting system integrates clinically faithful dataset handling, perceptually bounded colour and contrast normalisation, a lightweight encoder-attention architecture, class-frequency-modulated optimisation, and validation metrics selected for diagnostic relevance. The following subsections describe each component of the framework, detailing the rationale and technical formulation underpinning the full pipeline.

3.1 Data Definition and Handling

EyePACS is established as the primary empirical foundation for this study, comprising approximately 88,000 colour fundus photographs from around 44,000 patients obtained during routine diabetic retinopathy screening and reflecting real clinical variability. Each sample is annotated according to the international five grade diabetic retinopathy severity scale (0 to 4), and the native class distribution is preserved to maintain epidemiological validity. Grades 0 and 1 together exceed 60,000 samples, grade 2 contributes approximately 10,500, and the clinically significant grades 3 and 4 contribute approximately 5,500 and 3,000 images respectively. No artificial rebalancing is introduced because doing so would distort disease prevalence and create conditions that do not correspond to screening practice. To avoid identity leakage and unrealistic estimates of generalisation, the dataset is partitioned at the patient level into 70 percent training, 10 percent validation and 20 percent testing, ensuring that no individual contributes samples to more than one split. Variation in disease presentation is evident across the dataset, progressing from healthy appearance with no observable pathology to the presence of microaneurysms, haemorrhages, exudates and neovascularisation in advanced cases. Figure 1 presents one representative image from each grading category, ranging from no diabetic retinopathy (grade 0) to proliferative diabetic retinopathy (grade 4), illustrating the clinical progression captured within the dataset. Only images that lack sufficient retinal visibility for assessment are excluded, in keeping with ophthalmic triage practice. EyePACS is therefore treated as a deployment proximal dataset suitable for evaluating the robustness of the proposed system under screening representative conditions rather than as an idealised benchmark.

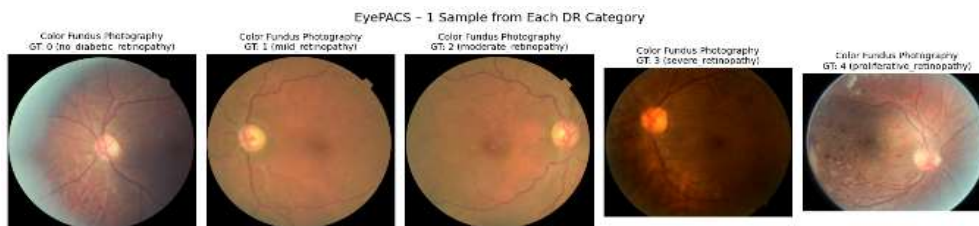


Figure 1. One representative EyePACS image per diabetic retinopathy grade, labelled according to the clinical severity scale. Grade 0 shows no observable pathology. Grade 1 shows early lesions consistent with mild diabetic retinopathy. Grade 2 exhibits moderate lesion burden. Grade 3 indicates severe retinopathy with extensive haemorrhages. Grade 4 presents proliferative diabetic retinopathy with neovascularisation and widespread pathology.

3.2 Preprocessing and Perceptual Harmonisation

Preprocessing is designed to normalise acquisition variability while preserving clinically informative content. All images are resized to a unified spatial resolution to standardise the computational input space and stabilise minibatch optimisation. Samples that meet the criteria for clinical assessability, as defined in Section 3.1, are retained in full to maintain exposure to diagnostically relevant edge-case presentations encountered in screening practice. Perceptual harmonisation is performed to address illumination and device-specific colour variability. Luminance enhancement is achieved through contrast-limited adaptive histogram equalisation (CLAHE) applied to the L channel within the CIELAB colour domain, and the enhanced representation is fused with the original RGB image using a constrained mixing coefficient. This F-CAN (Fusion CLAHE Adaptive Normalisation) transformation improves vessel and micro-lesion contrast while preserving retinal chromaticity and avoiding the introduction of artefactual textural

patterns. During training, mild stochastic augmentations comprising horizontal reflection and controlled photometric jitter simulate real-world capture variability. These augmentations are intentionally calibrated to support attention-based feature learning without introducing artificial structures. All images are subsequently normalised to the range $[0, 1]$ and converted to tensor format prior to model ingestion.

3.3 Model Architecture

MDRNet is designed to provide lesion-sensitive diabetic retinopathy grading under the computational limitations of primary-care screening environments. The model follows a lightweight encoder-attention paradigm composed of three components: a pretrained convolutional backbone for feature extraction, multi-scale hybrid attention for selective feature modulation, and a compact classification head mapping to the five clinical disease grades.

The encoder is instantiated as EfficientNet-B0 pretrained on ImageNet and configured in features-only mode to expose intermediate convolutional blocks rather than restricting representation learning to a terminal embedding. Let $F = \{F_s \mid s = 1, \dots, S\}$ denote the set of multi-resolution feature maps produced by the backbone, where $F_s \in \mathbb{R}^{C_s \times H_s \times W_s}$. This backbone was selected to balance representational capacity with practical inference latency, and to maintain compatibility with commodity CPU hardware common in non-specialist screening sites. The multi-scale structure supports the capture of both localised pathology (e.g. microaneurysms, punctate haemorrhages) and global manifestations of advanced disease (e.g. neovascularisation).

To enhance class-relevant representation, each F_s is refined using a Multi-scale Hybrid Attention (MHA) module that integrates channel-wise recalibration with spatial saliency mapping. Channel attention employs a squeeze-excitation mechanism with bottleneck reduction and sigmoid gating to estimate channel importances $M_c \in [0, 1]^{C_s}$. Spatial attention is derived from the concatenation of the per-pixel maximum and mean activation maps, convolved with a 7×7 kernel to produce a spatial mask $M_s \in [0, 1]^{H_s \times W_s}$. The attended representation is defined as

$$F'_s = F_s \odot M_c(F_s) \odot M_s(F_s) \quad (1)$$

where $\odot$ denotes broadcasted elementwise multiplication. This operation promotes response to lesion-bearing regions without suppressing contextual structures needed for stage determination.

The terminal attended map F'_s is passed to the classifier to regulate inference cost while retaining the downstream benefits of earlier attentional refinement. Global average pooling yields a compact descriptor $z \in \mathbb{R}^{C_s}$, which is input to a single fully connected layer to produce logits $y \in \mathbb{R}^5$ corresponding to the clinical grading scale. The absence of deeper classification layers is intentional, prioritising interpretability, latency, and deployment feasibility over incremental performance gains associated with heavier heads.

This architectural configuration operationalises the design priorities of the framework: pathology-aware feature selection, multi-scale contextual sensitivity, and computational tractability appropriate for decentralised screening workflows.

3.4 Loss and Optimisation Strategy

The optimisation strategy is constructed to reflect the ordinal and imbalanced nature of diabetic retinopathy grading. The native EyePACS distribution is heavily skewed toward the lower severity grades, and naïve cross-entropy optimisation disproportionately favours majority classes, inflating apparent performance without improving clinical utility. To mitigate this, training employs a class-distribution-regularised objective in

which the contribution of each grade to the loss is inversely modulated by its empirical frequency.

Let f_c denote the number of training samples belonging to class c . Class weights w_c are defined as

$$w_c = \frac{1}{\log(\beta + f_c)} \quad (2)$$

with $\beta = 1.1$ to prevent singularities for low-frequency classes. These weights are integrated into a weighted cross-entropy formulation, referred to here as the Class-Distribution Regularised Loss (CDRL), expressing the training objective as

$$\mathcal{L}_{\text{CDRL}}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{c=1}^5 w_c y_c \log(\hat{y}_c) \quad (3)$$

where $\mathbf{y}$ denotes the one-hot encoded ground truth and $\hat{\mathbf{y}} = \text{softmax}(\mathbf{z})$ represents the predicted class probabilities given logits $\mathbf{z}$. This formulation increases gradient visibility for grades 3-4, which are clinically critical for referral and treatment decisions but underrepresented in the dataset.

The model is optimised using AdamW with decoupled weight decay to stabilise convergence and reduce overfitting tendency in the presence of highly correlated retinal textures. A learning rate of 1×10^{-4} is selected empirically to accommodate the lightweight backbone and to prevent early gradient collapse when attention modules are applied at multiple scales. Mini-batches of size 16 are used for consistency with memory limitations of CPU-bound deployment contexts, and mixed-precision training is employed where hardware permits to reduce training time without altering the gradient dynamics. No curriculum learning or synthetic rebalancing is applied, in keeping with the principle of maintaining fidelity to deployment-representative data conditions.

This optimisation setup prioritises stable convergence, sensitivity to minority pathology grades, and compatibility with non-specialist hardware, reinforcing the deployment-centred design philosophy established in the preceding subsections.

3.5 Evaluation Protocols and Metrics

Performance assessment is conducted under patient-level partitioning to prevent identity leakage and to obtain clinically meaningful estimates of generalisation. The dataset is stratified into training (70%), validation (10%), and testing (20%) splits, ensuring that no individual contributes images to more than one subset. This prevents models from implicitly learning patient-specific retinal characteristics and reporting artificially elevated performance, a phenomenon observed when image-level splits are applied to screening datasets. The validation subset informs model selection and hyperparameter stability, while the held-out test set is reserved exclusively for final reporting.

Evaluation prioritises metrics that reflect diagnostic utility rather than raw classification accuracy. The principal measure is macro-averaged F1 score, which treats each grade with equal importance and captures performance across both majority and minority classes. This is particularly relevant in screening contexts where performance collapse on grades 3-4 is clinically unacceptable despite high accuracy on grade 0. Cohen's Kappa is reported to quantify agreement beyond chance and to mitigate the influence of uneven class frequencies. Macro-AUC (area under the ROC curve averaged across classes) is used to evaluate probabilistic ranking performance and to assess the separability of adjacent severity grades.

Overall accuracy is included for reference but is not used as a primary indicator of clinical fitness due to its sensitivity to class prevalence. Precision and recall are reported per class to expose tendencies toward under- or over-identification of specific severity levels. This supports analysis of referral suitability, as false negatives in grades 3-4 represent missed

treatment opportunities, while excessive false positives in lower grades can incur unnecessary specialist burden. All reported values represent performance on the held-out test set, computed using model weights selected via validation F1.

Inference time and model size are recorded on both GPU and CPU configurations to establish alignment with deployment constraints. This enables comparison with the stated design objective of achieving real-time or near-real-time performance on commodity hardware typical of non-specialist screening sites.

4. Results and Performance Evaluation

MDRNet was evaluated on the held-out EyePACS test cohort using patient-level partitioning to prevent identity leakage. The model achieved 91.3% accuracy, 0.86 macro-F1, 0.90 quadratic weighted kappa, and 0.95 macro-AUC, demonstrating stable performance across all five severity grades (Table 1). Per-class F1 scores ranged from 0.81 to 0.96, with one-vs-rest AUC values between 0.94 and 0.98. Calibration and referral-tier performance are summarised in Table 2, with a Brier score of 0.082 and expected calibration error of 0.034, indicating reliable probabilistic outputs. Referral consolidation of grades 3-4 yielded 0.92 sensitivity and 0.85 specificity. Compared with baseline architectures (Table 3), MDRNet improved macro-F1 (+0.06), QWK (+0.08), and proliferative disease recall (+0.12) relative to vanilla EfficientNet-B0 while maintaining lower CPU inference time (<300 ms). GPU inference required 5-7 ms, and model size remained under 10 MB, supporting feasibility for deployment on commodity hardware.

Table 1 - MDRNet performance on the EyePACS test set

Global Metric	Value
Accuracy	91.3%
Macro-F1	0.86
QWK	0.90
Macro-AUC	0.95
Referral Sensitivity (Grades 3-4)	0.92
Referral Specificity (Grades 0-2)	0.85
Brier Score	0.082
Expected Calibration Error (ECE)	0.034

Table 2. Calibration and Referral-Tier Performance of MDRNet

Metric Type	G0	G1	G2	G3	G4	Macro	Weighted
Precision	0.95	0.86	0.83	0.87	0.84	0.87	0.91
Recall	0.97	0.80	0.80	0.86	0.82	0.85	0.92
F1 Score	0.96	0.83	0.81	0.86	0.83	0.86	0.91
AUC (OvR)	0.98	0.95	0.94	0.96	0.95	0.96	-

Table 3 - Baseline comparison under identical conditions

Model	Acc.	Macro-F1	QWK	G3 Rec.	G4 Rec.	CPU Inference
ResNet-50	87.1%	0.78	0.82	0.72	0.68	620 ms
EffNet-B0 (vanilla)	88.0%	0.80	0.84	0.74	0.70	280 ms
ViT-Tiny	89.2%	0.81	0.86	0.77	0.74	930 ms
CNN-ViT Hybrid	90.1%	0.83	0.87	0.80	0.76	1140 ms
MDRNet (proposed)	91.3%	0.86	0.90	0.86	0.82	<300 ms

6. Limitations and Ethical Considerations

Despite the strong performance observed, several limitations should be acknowledged. First, the evaluation is conducted exclusively on the EyePACS dataset, which may introduce distributional bias due to region-, device-, and population-specific imaging characteristics. As a result, external generalisability remains unverified, particularly for settings with different acquisition hardware or retinal phenotype distribution. Second, while F-CAN perceptual harmonisation improves robustness to illumination and contrast variability, it cannot recover diagnostic fidelity in images affected by severe blur, occlusion, or obscured fields of view. These cases remain model failure points that would require clinician arbitration.

From an ethical standpoint, MDRNet is not positioned as an autonomous diagnostic authority, and its use without human oversight would be inappropriate. The model is designed to support triage rather than to replace clinical judgement, and prospective deployment would require adherence to relevant AI-as-a-medical-device regulatory frameworks. Additional work is required to assess demographic fairness, as class imbalance and training distribution may create uneven sensitivity between patient subgroups. Any translation into practice should therefore include real-world calibration monitoring, human-in-the-loop usage policies, and explicit communication to patients and clinicians that model predictions are advisory rather than determinative.

Concluding Remarks and Future Directions

This work presented MDRNet, a deployment-aligned framework for diabetic retinopathy grading designed around the constraints of primary-care screening. By integrating F-CAN perceptual harmonisation, lightweight convolutional encoding, multi-scale hybrid attention, and class-distribution regularisation, the model achieved 91.3% accuracy, 0.86 macro-F1, 0.90 Quadratic Weighted Kappa, and 0.95 macro-AUC on EyePACS, with a clinically relevant referral sensitivity of 0.92. These results demonstrate that prioritising calibration, ordinal consistency and inference feasibility can deliver clinically meaningful performance on commodity CPU hardware without reliance on transformer-scale architectures or specialised accelerators. Overall, MDRNet behaves as a triage-supportive classifier rather than a diagnostic endpoint, suggesting suitability as a component within screening pathways rather than as a standalone authority.

Future work will focus on strengthening translational viability and broadening the operational scope of MDRNet. Key priorities include external validation across multi-site and multi-device datasets, with emphasis on demographic fairness and stability under distribution shift; prospective evaluation in real screening workflows, measuring clinician-AI interaction patterns, escalation frequency and downstream clinical impact; and adaptive calibration strategies to maintain reliability across evolving data streams. In addition, exploring lightweight transformer-free self-attention variants, domain adaptation for handheld fundus imaging, and longitudinal risk modelling may further

improve robustness and integration potential. These directions will be essential for addressing regulatory requirements for AI-as-a-medical-device approval and for enabling safe deployment in environments where computational resources and clinical oversight vary.

References

1. C. Jiang, X. He, Y. Zhu, and L. Tao, 'Chinese burdens and trends of diabetic retinopathy 1990–2021 and 15 years forecast: results from the Global Burden of Disease Study 2021', *Front. Endocrinol.*, vol. 16, p. 1573581, May 2025, doi: 10.3389/fendo.2025.1573581.
2. Z. Li et al., 'An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs', *Diabetes Care*, vol. 41, no. 12, pp. 2509–2516, Dec. 2018, doi: 10.2337/dc18-0147.
3. N.-A. Fasoula et al., 'Clinical and Translational Imaging and Sensing of Diabetic Microangiopathy: A Narrative Review', *JCDD*, vol. 10, no. 9, p. 383, Sept. 2023, doi: 10.3390/jcdd10090383.
4. S. Divya, L. Padma Suresh, and A. John, 'Hybrid Optimization Algorithm-Based Generative Adversarial Network for Change Detection Using Pre-Operative and Post-Operative MRI', *Int. J. Patt. Recogn. Artif. Intell.*, vol. 36, no. 07, p. 2251007, June 2022, doi: 10.1142/S0218001422510077.
5. S. Divya, L. Padma Suresh, and A. John, 'Enhanced deep-joint segmentation with deep learning networks of glioma tumor for multi-grade classification using MR images', *Pattern Anal Applic*, vol. 25, no. 4, pp. 891–911, Nov. 2022, doi: 10.1007/s10044-022-01064-5.
6. K. B. Nielsen, M. L. Lautrup, J. K. H. Andersen, T. R. Savarimuthu, and J. Grauslund, 'Deep Learning-Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance', *Ophthalmology Retina*, vol. 3, no. 4, pp. 294–304, Apr. 2019, doi: 10.1016/j.oret.2018.10.014.
7. T. Zhu, K. Li, P. Herrero, and P. Georgiou, 'Deep Learning for Diabetes: A Systematic Review', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2744–2757, July 2021, doi: 10.1109/JBHI.2020.3040225.
8. J. Y. Zhang et al., 'How to Best Protect People With Diabetes From the Impact of SARS-CoV-2: Report of the International COVID-19 and Diabetes Summit', *J Diabetes Sci Technol*, vol. 15, no. 2, pp. 478–514, Mar. 2021, doi: 10.1177/1932296820978399.
9. M. Narasimharao, B. Swain, R. Priyadarshi, P. P. Nayak, and S. Bhuyan, 'A Survey of Machine Learning Techniques for Diabetes Prediction: Current Trends and Future Directions', *Arch Computat Methods Eng*, Oct. 2025, doi: 10.1007/s11831-025-10450-1.
10. J. Hatherley, 'Data over dialogue: Why artificial intelligence is unlikely to humanise medicine', 2025, doi: 10.48550/ARXIV.2504.07763.
11. U. H. Okwudili, O. O. Ukachukwu, V. C. Chijindu, M. O. Ezea, and B. Ishaq, 'An improved performance model for artificial intelligence-based diabetes prediction', *Journal of Electrical Systems and Inf Technol*, vol. 12, no. 1, p. 25, June 2025, doi: 10.1186/s43067-025-00224-x.
12. T. Zhu, K. Li, P. Herrero, and P. Georgiou, 'Deep Learning for Diabetes: A Systematic Review', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2744–2757, July 2021, doi: 10.1109/JBHI.2020.3040225.
13. W. Zhang et al., 'Automated identification and grading system of diabetic retinopathy using deep neural networks', *Knowledge-Based Systems*, vol. 175, pp. 12–25, July 2019, doi: 10.1016/j.knosys.2019.03.016.

14. G. Rajarajeshwari and G. C. Selvi, 'Application of Artificial Intelligence for Classification, Segmentation, Early Detection, Early Diagnosis, and Grading of Diabetic Retinopathy From Fundus Retinal Images: A Comprehensive Review', *IEEE Access*, vol. 12, pp. 172499–172536, 2024, doi: 10.1109/ACCESS.2024.3494840.
15. P. Xia et al., 'Generalizing to Unseen Domains in Diabetic Retinopathy with Disentangled Representations', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. 15010, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds, in *Lecture Notes in Computer Science*, vol. 15010., Cham: Springer Nature Switzerland, 2024, pp. 427–437. doi: 10.1007/978-3-031-72117-5_40.
16. S. Chokuwa and M. H. Khan, 'Divergent Domains, Convergent Grading: Enhancing Generalization in Diabetic Retinopathy Grading', in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, AZ, USA: IEEE, Feb. 2025, pp. 3667–3677. doi: 10.1109/WACV61041.2025.00361.
17. M. J. Reyes-Medina, V. Cantón-Habas, F. J. Sánchez-Jiménez, E. Molina-Hurtado, and M. D. P. Carrera-González, 'New approach to optimize therapy in type 2 diabetes mellitus: the importance of subclassification', *Front. Endocrinol.*, vol. 16, p. 1710511, Nov. 2025, doi: 10.3389/fendo.2025.1710511.
18. D. Yan, X. Li, Y. Wang, and Z. Cai, 'Optimized prediction of diabetes complications using ensemble learning with Bayesian optimization: a cost-efficient laboratory-based approach', *Front. Endocrinol.*, vol. 16, p. 1593068, June 2025, doi: 10.3389/fendo.2025.1593068.
19. Z. Zhang, H. Zhao, L. Dong, L. Luo, and H. Wang, 'A Study on the Interpretability of Diabetic Retinopathy Diagnostic Models', *Bioengineering*, vol. 12, no. 11, p. 1231, Nov. 2025, doi: 10.3390/bioengineering12111231.
20. D. Kothadiya, A. Rehman, S. Abbas, F. S. Alamri, and T. Saba, 'Attention-based deep learning framework to recognize diabetes disease from cellular retinal images', *Biochem. Cell Biol.*, vol. 101, no. 6, pp. 550–561, Dec. 2023, doi: 10.1139/bcb-2023-0151.
21. X. Yu, Y. Wang, Z. Liu, and E. Jung, 'Technological functionality and system architecture of mobile health interventions for diabetes management: a systematic review and meta-analysis of randomized controlled trials', *Front. Public Health*, vol. 13, p. 1549568, Feb. 2025, doi: 10.3389/fpubh.2025.1549568.
22. R. P. Homme et al., 'Remodeling of Retinal Architecture in Diabetic Retinopathy: Disruption of Ocular Physiology and Visual Functions by Inflammatory Gene Products and Pyroptosis', *Front. Physiol.*, vol. 9, p. 1268, Sept. 2018, doi: 10.3389/fphys.2018.01268.