# A Study of Diabetic Retinopathy Grading Using Robust Deep Learning

*Gunjan Mittal Roy[1], Ajay Kumar [2]*
[1]Lincoln University College Malaysia, [2]IILM University, Greater Noida, India
[1]Email ID: gunjanmittal2013@gmail.com; [2]Email ID: ajay.phdcse@gmail.com

**Abstract:** Diabetic Retinopathy (DR) is a leading cause of preventable blindness worldwide and presents a growing burden on global healthcare systems. While deep learning-based automated DR grading systems have achieved expert-level performance on benchmark datasets, their real-world deployment remains constrained by class imbalance, annotation noise, heterogeneous imaging conditions, and domain variability. This study presents a comprehensive review and structured analysis of robust deep learning strategies for DR grading under realistic clinical environments. Dataset characteristics, architectural developments, imbalance-aware optimization techniques, noise-resilient training methods, and domain adaptation approaches are systematically examined. Additionally, evaluation protocols, clinical validation requirements, and ethical considerations for trustworthy AI deployment are discussed. The study proposes a unified robustness-driven framework to enhance generalization, reliability, and scalability of DR grading systems in real-world screening programs.
(a) Problem statement/motivation of the article (b) solution (c) significant findings (d) applications

**Keywords**: Diabetic Retinopathy; Deep Learning; Imbalanced Data; Noisy Labels; Domain Adaptation; Clinical AI

## Introduction

Diabetic Retinopathy (DR) is a progressive retinal disorder resulting from prolonged hyperglycaemia and remains a major cause of visual impairment globally [1], [2]. Early detection and severity grading are critical to prevent irreversible blindness. Traditional screening relies on manual examination of fundus photographs by ophthalmologists. Although clinically effective, this process is labor-intensive and subject to inter-grader variability [5]. Deep learning, particularly Convolutional Neural Networks (CNNs), has significantly improved automated DR detection and grading [1], [3]. Regulatory-approved AI-based diagnostic systems further demonstrate translational potential [4]. However, most existing models are trained on curated datasets and may not generalize to real-world heterogeneous environments [17]. Therefore, robust learning strategies are essential for scalable deployment.

## CLINICAL BACKGROUND AND GRADING PROTOCOL

DR is categorized into five severity levels: No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and Proliferative DR (PDR) [7]. Lesion types such as microaneurysms, haemorrhages, exudates, and neovascularization determine grading.

Mathematically, DR grading is treated as a multi-class or ordinal classification task. Quadratic Weighted Kappa (QWK) is commonly used to reflect ordinal misclassification severity [5]. Inter-grader variability introduces annotation noise, which significantly affects supervised learning models [5].

**END-TO-END ROBUST DR GRADING FRAMEWORK**

The framework integrates preprocessing, imbalance mitigation, noise-resilient training, hybrid feature extraction, domain adaptation, and explainability modules.

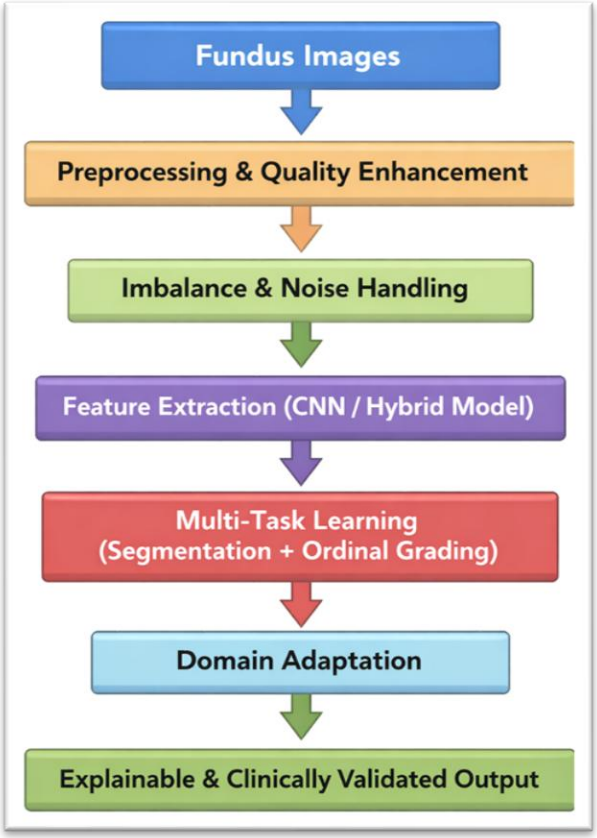


*Figure 1. End-to-End Robust Deep Learning Framework for DR Grading.*

Figure 1 illustrates the overall pipeline for automated analysis of retinal fundus images. The images are first preprocessed and enhanced, followed by handling class imbalance and noise, and then deep feature extraction using CNN or hybrid models. These features are used in a multi-task learning framework for lesion segmentation and disease grading, followed by domain adaptation to improve generalization, producing explainable and clinically validated diagnostic outputs.

**DATASET CHARACTERISTICS AND ROBUSTNESS CHALLENGES**

Publicly available datasets have played a central role in advancing automated diabetic retinopathy (DR) grading research. However, these datasets differ substantially in size, annotation protocol, acquisition settings, and population diversity. EyePACS [8], one of the largest screening datasets, reflects real-world variability in illumination, focus quality, and camera devices. Nevertheless, it exhibits severe class imbalance and annotation inconsistencies due to multi-grader labeling. Messidor [9], in contrast, provides high-quality curated images but lacks large-scale demographic diversity. APTOS 2019 [10] introduces practical variability with moderate imbalance, while IDRiD [11] includes pixel-level lesion annotations suitable for segmentation tasks. The DDR dataset [6] further incorporates real-world multi-center clinical images, enabling broader deployment-oriented evaluation.

Despite their complementary strengths, these datasets do not fully capture longitudinal disease progression, cross-device variability, and extreme minority-class representation. Consequently, models trained on a single dataset often exhibit substantial performance degradation during cross-dataset validation [17].

### A. Class Imbalance in DR Grading

Class imbalance remains one of the most critical challenges in DR screening systems. Severe NPDR and PDR cases are significantly underrepresented relative to No DR or Mild DR categories. Standard cross-entropy loss biases optimization toward majority classes, reducing sensitivity for clinically urgent cases.

Algorithm-level imbalance mitigation strategies have shown promising results. Focal loss [12] introduces a modulating factor that down-weights easy examples and focuses learning on hard or minority samples. Class-balanced loss [13] reweights classes based on the effective number of samples, improving gradient contribution from underrepresented categories. Additionally, weighted sampling and data augmentation techniques can further stabilize minority-class learning.

However, imbalance-aware learning must be carefully tuned to avoid over-amplifying noisy minority samples. A hybrid approach combining balanced loss functions with noise-resilient training often yields superior results.

### B. Noisy Annotations and Inter-Grader Variability

DR grading inherently involves subjective interpretation. Krause et al. [5] demonstrated that adjudicated grading significantly improves reference reliability compared to single-grader annotations. Label noise, whether symmetric or asymmetric, may cause deep neural networks to memorize corrupted labels during prolonged training. To address this issue, noise-robust optimization methods have been introduced. Loss correction techniques [14] estimate label transition probabilities to adjust training targets. Generalized cross-entropy loss (GCE) [15] balances between mean absolute error and cross-entropy, reducing sensitivity to mislabelled samples. Co-teaching [16] trains two networks simultaneously, allowing each model to select small-loss samples for peer updating, thereby filtering noisy data.

These methods prevent performance collapse under high noise rates and enhance convergence stability in realistic screening datasets.

### C. Domain Shift and Generalization Constraints

Domain shift arises from variations in imaging devices, resolution, colour distribution, and patient demographics. Models trained on one dataset may fail to generalize to external cohorts due to covariate

shift. Voets et al. [17] demonstrated reproducibility gaps when models were evaluated across independent datasets. Domain adaptation methods aim to reduce feature distribution discrepancies between source and target domains. Feature alignment techniques, adversarial training strategies, and cross-dataset validation protocols improve deployment readiness. Nevertheless, domain adaptation requires representative target-domain samples and may introduce optimization instability if not carefully designed. Robust DR grading therefore demands integrated solutions that simultaneously address imbalance, noise, and domain variability.

## DEEP LEARNING ARCHITECTURES FOR DR GRADING

Deep learning architectures for DR grading have evolved significantly over the past decade. Early implementations employed CNNs inspired by AlexNet and VGG-like architectures [3], [18]. With the introduction of residual learning, ResNet [18] enabled deeper networks with improved gradient propagation, significantly enhancing medical image feature extraction. DenseNet [19] further improved feature reuse through dense connectivity, reducing parameter redundancy while preserving discriminative power. Transfer learning has been widely adopted in DR grading due to limited availability of labeled medical images. Tajbakhsh et al. [20] demonstrated that fine-tuning pre-trained networks significantly improves performance compared to training from scratch in medical imaging applications. Pre-trained CNN backbones extract hierarchical retinal features, including vascular patterns, lesion textures, and structural abnormalities.

More recent studies incorporate attention mechanisms to explicitly model lesion-specific regions. Attention modules dynamically assign higher weights to clinically relevant areas such as microaneurysms and hemorrhages, thereby improving interpretability and localization accuracy [22]. Multi-task learning architectures further enhance robustness by jointly optimizing segmentation and classification objectives. For instance, lesion segmentation branches guide the classification head to focus on pathological regions [6].

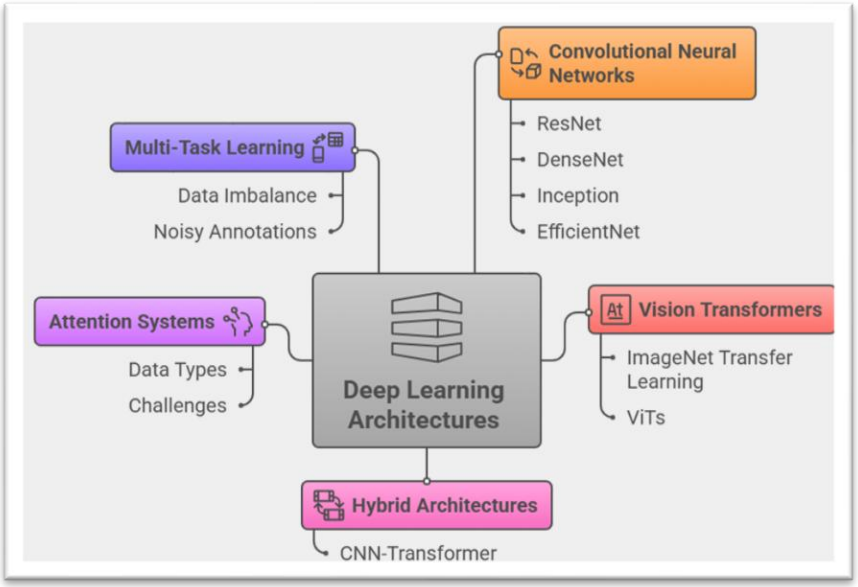


*Figure 2. Deep Learning Architectures for Diabetic Retinopathy Grading*

Hybrid CNN–Transformer architectures extend these capabilities by integrating convolutional local feature extraction with global self-attention modeling. Such models capture long-range dependencies across the retinal field, which is particularly beneficial in detecting distributed lesions. Systematic reviews indicate that ensemble and hybrid architectures achieve improved robustness under heterogeneous screening conditions [21], [23]. Despite architectural advancements, increasing model complexity must be balanced against computational efficiency, especially for deployment in resource-constrained clinical settings.

*Table 1. COMPARATIVE ANALYSIS OF ROBUSTNESS STRATEGIES*

| Strategy | Objective | Methods | Advantages | Limitations |
|---|---|---|---|---|
| Imbalance-Aware Learning | Improve minority sensitivity | Focal Loss [12], Class-Balanced Loss [13] | Better severe DR detection | Sensitive to noisy rare samples |
| Noise-Resilient Learning | Handle label corruption | Loss Correction [14], GCE [15], Co-teaching [16] | Stable training under noise | Increased computational complexity |
| Domain Adaptation | Improve cross-dataset generalization | Feature alignment, cross-dataset validation [17] | Deployment readiness | Requires multi-domain data |

**EVALUATION AND CLINICAL VALIDATION**

Rigorous evaluation is critical for assessing deployment readiness of DR grading systems. Commonly reported metrics include accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and Quadratic Weighted Kappa (QWK) [5]. For screening programs, sensitivity for referable DR is prioritized to minimize false negatives, while maintaining acceptable specificity to avoid unnecessary referrals. However, retrospective evaluation on a single dataset is insufficient to establish generalizability. Multi-center validation across diverse demographic cohorts significantly strengthens reliability claims [2], [4]. Cross-dataset testing reveals the extent of domain shift and provides realistic estimates of clinical performance [17]. Calibration analysis is another important but often overlooked aspect. Well-calibrated probability outputs are essential for risk stratification and clinical decision support. Miscalibrated models may lead to inappropriate referral thresholds. Explainability mechanisms such as saliency maps and attention visualization enhance transparency and clinician trust. Nevertheless, visual explanations must be validated to ensure they correlate with true pathological regions.

Regulatory approval further requires documentation of model robustness, failure cases, bias analysis, and reproducibility. Ethical AI deployment demands fairness assessment across demographic subgroups to prevent algorithmic discrimination. Consequently, evaluation protocols must extend beyond performance metrics to include reliability, interpretability, and safety considerations.

**LIMITATIONS**

Although robust strategies improve generalization, extreme imbalance and severe annotation corruption may still bias training. Domain adaptation requires access to representative multi-center datasets, which

may be restricted due to privacy constraints. Most studies remain retrospective; prospective longitudinal validation is limited. Explainability methods also lack standardized clinical interpretation protocols.

*Table 2. Strategies for Handling Imbalance, Label Noise, and Domain Shift in Deep Learning-Based DR Detection.*

| Strategy Category | Primary Objective | Representative Methods | Advantages | Limitations | Suitable Use Case |
|---|---|---|---|---|---|
| Imbalance-Aware Learning | Improve minority-class sensitivity (e.g., Severe NPDR, PDR) | Focal Loss [12], Class-Balanced Loss [13], Weighted Sampling | Enhances detection of clinically critical stages; simple integration into training pipeline | May over-amplify rare noisy samples; does not address label corruption | Large-scale screening datasets (e.g., EyePACS) with severe class imbalance |
| Noise-Resilient Learning | Mitigate impact of annotation errors and inter-grader variability | Loss Correction [14], Generalized Cross-Entropy [15], Co-Teaching [16] | Prevents memorization of corrupted labels; improves convergence stability | Requires careful hyperparameter tuning; may increase training complexity | Datasets with subjective grading or inconsistent annotations |
| Domain Adaptation & Generalization | Reduce performance degradation across imaging devices and populations | Feature Alignment, Adversarial Domain Adaptation, Cross-Dataset Validation [17] | Enhances cross-center generalization; supports real-world deployment | Requires access to multi-domain data; adaptation may be unstable | Multi-center deployment environments with heterogeneous imaging protocols |

## CONCLUSION AND FUTURE SCOPE

Robust deep learning significantly advances DR grading but requires integrated imbalance mitigation, noise handling, and domain generalization strategies for real-world deployment. Future work should focus on longitudinal datasets, fairness-aware learning, federated training frameworks, and foundation-scale medical vision models for scalable global screening.

## References


1. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama, 316(22), 2402-2410. doi:10.1001/jama.2016.17216
2. Ting, D. S. W., Cheung, C. Y. L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., ... & Wong, T. Y. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye

diseases using retinal images from multiethnic populations with diabetes. *Jama*, *318*(22), 2211-2223.
http://dx.doi.org/10.1001/jama.2017.18152

3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539
4. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle
5. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., ... & Webster, D. R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, *125*(8), 1264-1272. https://doi.org/10.1016/j.ophtha.2018.01.034
6. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., ... & Jia, W. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature communications*, *12*(1), 3242. https://doi.org/10.1038/s41467-021-23458-5
7. Grzybowski, A., Kanclerz, P., Tsubota, K., Lanca, C., & Saw, S. M. (2019). The epidemiology of myopia in school children worldwide. *Acta Ophthalmologica*, *97*. https://doi.org/10.1111/j.1755-3768.2019.5485
8. EyePACS Dataset, 2015.
9. Messidor Dataset, 2008–2012.
10. APTOS Dataset, 2019.
11. P. Porwal et al., “IDRiD challenge,” 2018.
12. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988). https://doi.org/10.48550/arXiv.1708.02002
13. Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268-9277). https://doi.org/10.48550/arXiv.1901.05555
14. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1944-1952).
15. Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, *31*. https://doi.org/10.48550/arXiv.1805.07836
16. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., ... & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, *31*. https://doi.org/10.48550/arXiv.1804.06872

17. Voets, M., Mollersen, K., & Bongo, L. A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, *14*(6), e0217541. https://doi.org/10.1371/journal.pone.0217541
18. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
19. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE transactions on medical imaging*, *35*(5), 1299-1312. https://doi.org/10.1109/TMI.2016.2535302
20. Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y., ... & Ting, D. S. (2019). Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health*, *1*(1), e35-e44. https://doi.org/10.1016/S2589-7500(19)30004-4
21. Ryu, J. S., Kang, H., Chu, Y., & Yang, S. (2025). Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, *15*(5), 809-830. https://doi.org/10.1007/s13534-025-00484-6
22. Mondal, S. S., Mandal, N., Singh, K. K., Singh, A., & Izonin, I. (2022). Edldr: An ensemble deep learning technique for detection and classification of diabetic retinopathy. *Diagnostics*, *13*(1), 124. https://doi.org/10.3390/diagnostics13010124