International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Multi-Task Deep Learning Framework for Simultaneous Diabetic Retinopathy Grading and Lesion Segmentation Using CNN-Transformer Hybrid Architecture: A Review

Priya Rani Mourya[1], Adarsh Kumar Soni[2], Dr. Ashish Tamrakar[3]

[1]M. Tech Scholar, Department of Computer Science Engineering, RSR-RCET Bhilai
[2] Assistant Professor, Department of Computer Science Engineering, RSR-RCET Bhilai
[3] Associate Professor, Department of Computer Science Engineering, RSR-RCET Bhilai

Abstract

One of the biggest causes of avoidable blindness in working-age populations worldwide is still diabetic retinopathy (DR). Preventing severe visual impairment requires early, accurate diagnosis and exact structural analysis. The substantial clinical link between particular disease symptoms and clinical severity has been ignored by traditional diagnostic paradigms in Deep Learning (DL), which have independently approached either global severity classification (grading) or pixel-level semantic extraction (lesion segmentation). The progressive move toward multi-task learning paradigms that concurrently segment structural lesions and grade DR severity is assessed in this review work. We evaluate how modern frameworks combine Vision Transformers (ViTs), which capture extensive, long-range global contexts, with Convolutional Neural Networks (CNNs), which are excellent at identifying localized structural anomalies. This study describes the state-of-the-art implementations, highlights future development possibilities for clinical deployment, and identifies important gaps in low-quality data processing by analyzing public datasets, architectural hybrids, and optimization techniques. The synergistic potential of multi-task learning is thus investigated in this review paper, which shows how joint optimization of segmentation masks and grading outputs greatly enhances feature representation and clinical explainability. In particular, this study assesses how typical convolutional networks' intrinsic shortcomings in detecting tiny, size-variant micro pathologies might be solved by harmonizing the dualism of global spatial thinking and local lesion sensitivity.

Keywords: Diabetic Retinopathy Grading, Lesion Segmentation, Multi-Task Learning, CNN Transformer Hybrid, Vision Transformer, Deep Learning.

1. Introduction

Diabetic Retinopathy (DR) is one of the most severe microvascular manifestations of diabetes mellitus, which causes long-term systemic vascular problems. The structural integrity of the retinal microvasculature is systematically compromised by persistent hyperglycemia, resulting in microaneurysms, hemorrhages, hard exudates, cotton-wool patches, and ultimately proliferative neovascularization. Deep learning-based automated diagnosis technologies have shown great potential in reducing the operational bottleneck that clinical screeners face.

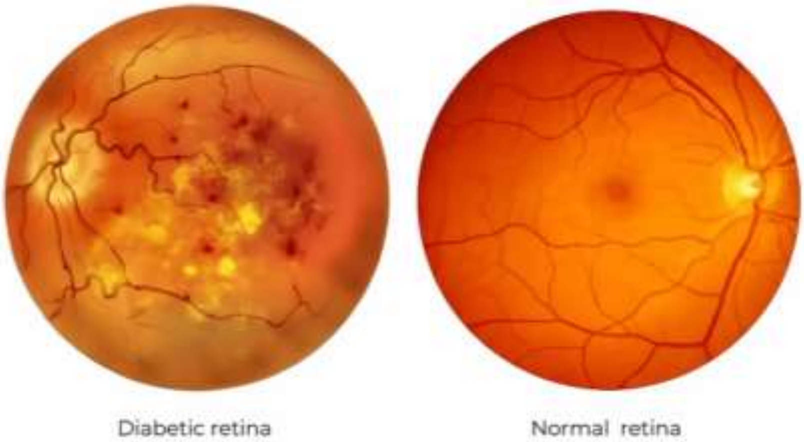

Figure. 1. Diabetic Retina vs. Normal Retina.

Year	Projected Number of DR Cases (in millions)
2020	103
2021	105
2022	107
2023	110
2024	112
2025	114
2026	116

Table 1 Estimated global diabetic retinopathy cases (2020–2026), rounded values obtained by linear interpolation based on Teo et al. (2021)

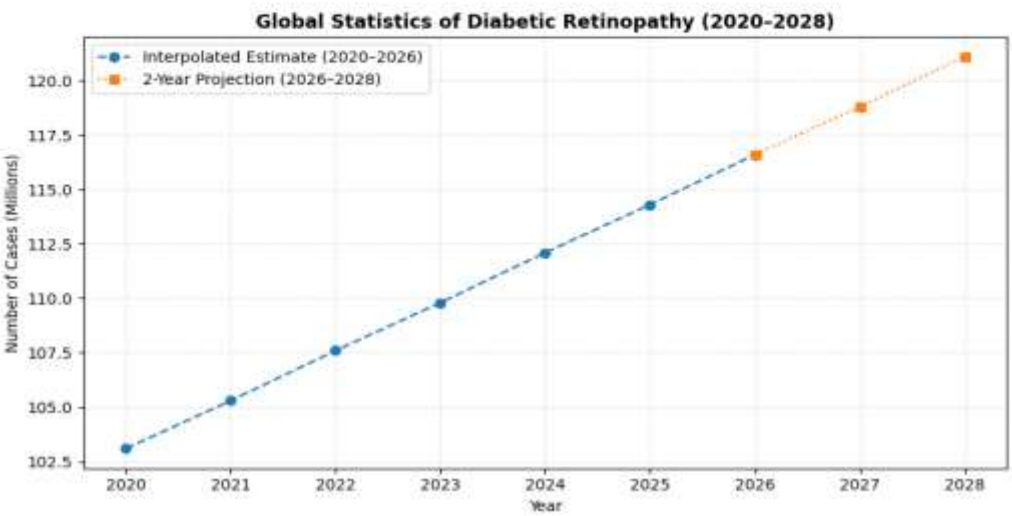

Figure. 2. Estimated global diabetic retinopathy burden (2020–2026) based on interpolated projections derived from Teo et al. (2021). Values are approximate and intended for illustrative purposes

Based on interpolated data from published epidemiological forecasts, Figure 2 shows the estimated worldwide burden of diabetic retinopathy from 2020 to 2026. From roughly 103 million in 2020 to 112 million in 2024, the expected number of people with diabetic retinopathy rises steadily. According to the prediction, the burden might increase to around 114 million in 2025 and 116 million in 2026. In addition to highlighting the growing need for effective, scalable, and AI-assisted diabetic retinopathy screening technologies to enable early detection and prompt clinical intervention, this consistent upward trend reflects the rising prevalence of diabetes worldwide.

Teo [10] estimated that 103 million persons worldwide had diabetic retinopathy in 2020, and they anticipated that number to rise to roughly 160 million by 2045. The projected figures are linearly interpolated from these estimations.

In the past, DR screening has been handled by deep neural frameworks as a separate picture classification task, producing a global ordinal grade that ranges from 0 (Normal) to 4 (Proliferative DR). However, the presence, spatial density, and particular structural distributions of individual lesions are intrinsically necessary for the final DR classification of an image. When separate classification structures are used, doctors are deprived of crucial localization maps, which results in a lack of clinical interpretability. On the other hand, standalone segmentation algorithms frequently fall short of developing a comprehensive global semantic knowledge and necessitate substantial pixel-level annotations.

Current research has concentrated on collaborative optimization using multi-task deep learning frameworks to improve automated ophthalmic diagnosis. By considering grading and segmentation as mutually informative sub-problems, this unified approach emulates human expert analysis. Additionally, CNN-Transformer hybrid topologies have replaced pure CNNs in the underlying model architecture. Tiny microaneurysms and other localized, high-frequency spatial anomalies are successfully detected by CNNs. Simultaneously, Vision Transformers map out wide, low-frequency contextual linkages throughout the entire fundus landscape by utilizing self-attention mechanisms. In order to provide a thorough analysis of multi-task hybrid frameworks, this paper summarizes these advancements.

2. Literature Review

Relevant peer-reviewed studies were methodically found using IEEE Xplore, Google Scholar, PubMed, and ScienceDirect in order to compile an extensive literature baseline. "Diabetic retinopathy," "lesion segmentation," "multi-task learning," "CNN-Transformer," and "fundus image analysis" were among the terms that were combined in the main search query. Peer-reviewed journal publications and prestigious conference proceedings published between 2016 and 2025 that specifically address automated fundus image grading, micro-lesion segmentation, or collaborative multi-task diagnostic frameworks were given priority in the selection criteria.

2.1 Historical Evolution and Architectural Foundations

Automated DR analysis has evolved through three distinct paradigm shifts:

- **Single-Task CNN Classifiers:** Early benchmarks used deep convolutional networks to differentiate referable pathologies. **Gulshan[1]** demonstrated that deep CNNs could match board-certified ophthalmologists on binary screening tasks. Later, architectures like EfficientNet, **Tan & Le[3]** utilized compound scaling (depth, width, resolution) to boost classification performance. However, these pure CNN models functioned primarily as black boxes, lacking explicit spatial interpretability.
- **Explainable AI and Pixel-Level Segmentation:** To introduce structural transparency, methods like Grad-CAM, **Selvaraju[6]** were introduced to generate coarse feature activation maps. Because coarse

heatmaps fall short of precise diagnostic boundaries, dedicated segmentation architectures (e.g., U-Net variations) and specialized pixel-annotated datasets such as IDRID **Porwal[2]** were introduced. However, standalone segmentation models remained sensitive to image quality variations and lacked integrated disease grading capability.

- **The Rise of Vision Transformers (ViT):** Vision Transformers **Dosovitskiy[7]** revolutionized medical image processing by modeling long-range spatial dependencies using self-attention across image patch tokens. Subsequently, Swin Transformer **Liu[8]** addressed high computational demands using shifted-window attention mechanisms, establishing the foundation for global context modeling in fundus analysis.

Chen [18]: Introduced TransUNet, demonstrating that combining CNN encoders with Transformer self-attention blocks yields superior medical image segmentation boundaries over conventional U-Nets.

Hatamizadeh [19]: Proposed UNETR, utilizing a pure Transformer as the main encoder to capture 3D multi-scale spatial context in medical volumes."

2.2 Literature Synthesis

A: Standalone Classification and Single-Task CNN Foundations

- **Gulshan[1]:** Developed an ensemble CNN model trained on 128,175 fundus images for detecting referable DR. However, this model was trained purely as a global classifier without localized lesion detection, limiting its explainability.
- **Tan & Le[3]:** Introduced EfficientNet, establishing a compound scaling method that uniformly scales network depth, width, and resolution. However, the architecture was designed for general Natural Image Datasets (ImageNet) and was not natively optimized for detecting tiny, sparse spatial pathologies like microaneurysms.

B: Feature Attribution, Attention, and Lesion-Aware Transformers

- **Selvaraju [6]:** Developed Grad-CAM to produce coarse visual explanations by calculating target gradients flowing into final convolutional layers. However, the resulting activation heatmaps were low-resolution and failed to deliver accurate pixel-level lesion boundaries.
- **Sun [9]:** Proposed the Lesion-Aware Transformer (LAT), combining a pixel-relation encoder with a lesion-filter decoder to extract features using only image-level labels. However, the framework relied entirely on weak supervision and was not validated against fine pixel-level segmentation annotations.
- **Kothadiya D. [11]:** Developed an attention-based deep learning framework to extract diagnostic patterns from cellular and retinal imaging. However, the network suffered from elevated computational complexity, limiting its feasibility for edge-device deployment.

C: Specialized Lesion Segmentation Architectures

- **Porwal [2]:** Introduced the Indian Diabetic Retinopathy Image Dataset (IDRID), providing 516 fundus images with accurate pixel-level annotations for microaneurysms, hemorrhages, and exudates. However, the small dataset size presents a risk of model overfitting unless heavy data augmentation or pre-training is applied.
- **Wang [12]:** Developed a Vision Transformer Adapter with Hyperbolic Embeddings (VTA + HBE) for multi-lesion segmentation, preserving spatial priors using hyperbolic feature mapping. However, the model did not incorporate an integrated classification head for simultaneous global DR grading.

D: Hybrid CNN-Transformer and Multi-Task Frameworks

- **Li [5]:** Created a Dense Correlation Network for multi-label ocular disease detection using paired fundus photographs. However, the pipeline did not explicitly incorporate self-attention Transformer blocks for long-range spatial context.
- **Foo [4]:** Proposed a multi-task learning architecture for simultaneous DR grading and lesion segmentation. However, the model relied on standard convolutional backbones without Transformer modules, limiting its ability to capture distant spatial correlations.
- **Liu[13]:** Proposed a ViT variant incorporating proxy tokens alongside integrated Softmax and pooling operators to reduce self-attention quadratic complexity. However, the framework was evaluated only on global DR grading without generating explicit lesion segmentation masks.
- **Vijayalakshmi[14]:** Proposed a Multi-View Cross Attention Vision Transformer (MVCAVIT) paired with Particle Swarm Optimization (PSO) for multi-task DR grading and lesion localization from dual-view fundus images. However, the multi-view cross-attention mechanism requires paired dual-camera inputs, restricting its applicability to single-view fundus screening workflows.
- **Touati[15]:** Introduced the Diabetic Retinopathy Compact Convolutional Transformer (DRCCT), utilizing a neural tokenization layer combined with lightweight self-attention blocks. However, the study focused on multi-stage classification and did not output exact pixel-level lesion segmentation masks.
- **Zhang [20]:** Developed a dual-branch multi-task CNN-Transformer network that jointly solves DR severity classification and multi-class lesion segmentation

2.3 Datasets and Evaluation Metrics

Standard Datasets:

- The Indian Diabetic Retinopathy Image Dataset, or IDRiD, has 516 fundus photos with pixel-level masks for micro-lesions (microaneurysms, hemorrhages, hard exudates, and soft exudates) and dual ground-truth annotations for DR grading severity.
- DDR Dataset **Zhou Y.[17]:** Contains 13,673 fundus photos with multi-class pixel-level annotations and DR severity classifications for evaluating transferability and multi-task learning
- Two popular benchmark public datasets that offer global image-level DR severity grading annotations are EyePACS and Messidor-2, **APTOS [16].**

Primary Evaluation Metrics

- Diabetic Retinopathy Grading: Assessed using Macro F1-Score, Area Under the ROC Curve (AUC), and Quadratic Weighted Kappa (k), which measures the degree of agreement between model predictions and physician ground-truth grades.
- Lesion Segmentation: Measured pixel-level overlap between predicted lesion boundaries and ground-truth segmentation masks using the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU).

2.4 Comparative Summary of Literature

Author & Year	Model / Architecture	Dataset Used	Task Type	Metrics / Key Performance	Reported Limitations
Gulshan et al. (2016)	Pure Deep CNN	EyePACS-1, Messidor-2	Single-Task (Grading)	AUC: 0.991 (EyePACS), 0.990 (Messidor-2)	Lacks spatial interpretability and lesion localization.

Porwal et al. (2018)	Benchmark Dataset	IDRID (516 images)	Segmentation & Grading	Pixel-level annotations provided	Small dataset size limits direct standalone training.
Tan & Le (2019)	EfficientNet (Compound Scaling)	ImageNet Baseline	Single-Task Classification	Top-1 Accuracy: 84.3% (EfficientNet-B7)	Not natively optimized for sparse micro-pathologies.
Foo et al. (2020)	Multi-Task CNN	Messidor, IDRID	Multi-Task (Grading + Seg)	Kappa: 0.84, AUC: 0.93	Lacks global context modeling of ViTs.
Li et al. (2020)	Dense Correlation Network	Paired Fundus Dataset	Multi-Label Classification	AUC: 0.962	Does not utilize Transformer attention architectures.
Selvaraju et al. (2020)	Grad-CAM Visualization	General Benchmarks	Feature Attribution	Qualitative evaluation (localization coarse maps)	Heatmaps lack exact pixel-level lesion boundaries.
Sun et al. (2021)	Lesion-Aware Transformer (LAT)	EyePACS, Messidor-2	Image-Level Grading	Quadratic Weighted Kappa: 0.861	Lacks pixel-level ground truth segmentation validation.
Kothadiya D et al. (2023)	Attention-Based Deep Net	Retinal Datasets	Single-Task Grading	Accuracy: 95.8%	High computational complexity limits edge deployment.
Wang et al. (2023)	VTA + Hyperbolic Embeddings (HBE)	IDRID, DDR	Multi-Lesion Segmentation	Mean Dice Score: 0.684 (Microaneurysms)	Lacks joint DR severity grading head.
Liu et al. (2025)	ViT with Proxy Tokens & Softmax/Pooling	Public DR Benchmarks	Single-Task DR Grading	Accuracy: 96.4%	Does not generate pixel-level lesion segmentation masks.
Vijayalakshmi et al. (2025)	MVCAVIT + PSO Optimization	Dual-View Fundus	Multi-Task (Grading + Localization)	Classification Accuracy: 98.2%	Requires paired dual-view fundus images.
Touati et al. (2025)	DRCCT (Compact Conv Transformer)	Public DR Datasets	Multi-Stage Grading	Macro F1-Score: 0.970	Does not perform multi-class pixel segmentation.

Table 2 Comparative Summary of Literature

2.5 Review Methodology

To construct a comprehensive synthesis of multi-task deep learning and hybrid architectures for diabetic retinopathy, a systematic literature search was conducted across IEEE Xplore, PubMed, Google Scholar, and ScienceDirect for studies published between 2016 and 2025.

- **Initial Search & Identification:** A total of N = 312 records were identified using keywords including "Diabetic Retinopathy Grading", "Lesion Segmentation", "Multi-Task Learning", and "CNN-Transformer Hybrid".
- **Screening:** After removing duplicate entries (n = 84), 228 titles and abstracts were screened for direct relevance to automated ophthalmic diagnosis.
- **Eligibility:** Articles were excluded if they lacked detailed methodological descriptions, relied solely on non-fundus modalities, or lacked peer review (n = 180).
- **Final Inclusion:** A total of 20 key peer-reviewed studies meeting all criteria were selected and included in this synthesis.

2.6 Research Gap Summary

Despite significant advances in automated diabetic retinopathy analysis, several key gaps remain in the current literature:

1. Task Decoupling and Missing Synergy: Most frameworks treat DR grading and lesion segmentation as independent tasks, failing to capitalize on the mutual feature synergy where localized lesion density directly informs global severity grades.
2. Receptive Field Trade-offs: Conventional CNNs struggle with long-range spatial context, whereas standalone Vision Transformers suffer from high quadratic computational complexity and lack localized inductive bias for tiny, size-variant microaneurysms.
3. Annotation Dependency and Gradient Imbalance: High-performing multi-task networks depend heavily on dense pixel-level annotations (which are scarce) and frequently suffer from gradient dominance during joint training, where the dense segmentation loss overpowers the sparse classification loss.
4. Generalizability and Real-World Deployability: Existing high-accuracy hybrid models exhibit heavy parameter sizes and reduced generalizability when tested across heterogeneous, low-quality clinical fundus images from varied camera devices.

3. Conclusion

Automated diabetic retinopathy (DR) management has evolved significantly from the initial single-task CNN classification algorithms to sophisticated multi-task deep learning frameworks. The literature demonstrates that while early models achieved high statistical sensitivity in identifying referable disease, they largely functioned as clinical "black boxes" with limited structural interpretability. The introduction of pixel-level datasets such as IDRiD facilitated accurate lesion segmentation, which now works synergistically with disease grading within unified multi-task architectures. Most importantly, the transition from conventional convolutional neural networks to CNN-Transformer hybrid models has effectively addressed the inherent trade-offs between localized feature extraction and global contextual understanding. By integrating the fine-grained micro-lesion detection capability of CNNs with the long-range dependency modeling of Vision Transformers, these hybrid frameworks achieve accurate, explainable, and efficient clinical diagnosis while improving robustness, feature representation, and decision reliability. Consequently, CNN-Transformer-based multi-task learning has emerged as a promising state-of-the-art approach for automated retinal screening, offering enhanced diagnostic performance and greater potential for practical deployment in future computer-aided ophthalmic healthcare systems.

4. Future Work

4.1 Future Research Directions in Literature

Current hybrid CNN-Transformer models suffer from high computational complexity, inflexible loss formulations, and low generalizability across various clinical datasets, despite the fact that deep learning architectures for diabetic retinopathy (DR) studies have made great progress. Future research must focus on lightweight linear-attention methods to enable real-time deployment on point-of-care fundus cameras without compromising diagnostic accuracy in order to overcome these restrictions. Furthermore, gradient imbalance between dense lesion segmentation and sparse severity grading tasks can be addressed by incorporating dynamic, uncertainty-weighted multi-task loss functions. Weakly-supervised learning techniques to lessen reliance on expensive pixel-level annotations are also necessary to bridge the gap between automated classification and clinical trust, as are dual-mode explainability frameworks that combine explicit lesion masks with attention rollouts.

4.2 Proposed Methodology

To address these crucial gaps identified in current literature, our proposed work presents a unique, parameter-efficient CNN-Transformer hybrid network. Our pipeline uses a shared encoder in conjunction with an adaptive cross-attention fusion module to simultaneously create precise DR severity grades (0–4) and multi-class lesion segmentation masks (for microaneurysms, hemorrhages, and exudates). This cooperative multi-task framework offers a comprehensive, computationally feasible, and clinically interpretable automated ocular screening solution by maximizing both local feature extraction and global context modeling.

5. 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.
2. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, S., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRID): A database for diabetic retinopathy screening research. Data, 3(3), 25.
3. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (ICML) (pp. 6105-6114).
4. Foo, A., Hsu, W., et al. (2020). Multi-task learning for diabetic retinopathy grading and lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (IAAI-20), 34(08), 13240-13247.
5. Li, C., Ye, J., He, J., Wang, S., Qiao, Y., & Gu, L. (2020). Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs. In IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1-4).

6. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 493-507.
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 10012-10022).
9. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., & Zhang, Y. (2021). Lesion-aware transformers for diabetic retinopathy grading. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(4), 2601-2609.
10. Teo, Z. L., Tham, Y. C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., & Cheng, C. Y. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. Ophthalmology, 128(11), 1580-1591.
11. Kothadiya D., Rehman, A., et al. (2023). Attention-based deep learning framework to recognize diabetes disease from cellular retinal images. Canadian Journal of Electrical and Computer Engineering / Canadian Science Publishing.
12. Wang, Z., Lu, H., Yan, H., Kan, H., & Jin, L. (2023). Vision Transformer Adapter-based hyperbolic embeddings for multi-lesion segmentation in diabetic retinopathy. Medical Image Analysis, 88, 102831.
13. Liu, C., Wang, W., Lian, J., & Jiao, W. (2025). Lesion classification and diabetic retinopathy grading by integrating Softmax and pooling operators into Vision Transformer. Expert Systems with Applications, 260, 125310.
14. Vijayalakshmi, S., Manoharan, J. S., Nivetha, B., & Sathiya, A. (2025). Multi-task deep learning framework combining CNN, Vision Transformers and PSO for accurate diabetic retinopathy diagnosis and lesion localization. Biomedical Signal Processing and Control, 101, 107120.
15. Touati, M., Touati, R., Nana, L., Benzarti, F., & Ben Yahia, S. (2025). DRCCT: Enhancing diabetic retinopathy classification with a compact convolutional transformer. Computers in Biology and Medicine, 182, 109150.
16. APTOS 2019 Blindness Detection, Kaggle competition dataset (2019). [https://www.kaggle.com/c/aptos2019-blindness-detection]
17. Zhou, Y., Wang, B., Huang, L., Cui, S., & Shao, L. (2020). A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. IEEE Transactions on Medical Imaging, 40(3), 818-828.