YOLOv8 + Vision Transformers (ViT): Hybrid Architecture for Fine-Grained DR Detection

K. Martin Sagayam[1,a], Shasi Kant Gupta [2], Sai Kiran Oruganti[3]

[1] Postdoctoral Researcher, Lincoln University College, Malaysia; [a] Division of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India

[2] Adjunct Research Faculty, Lincoln University College, Malaysia

Adjunct Research Faculty, Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, India;

[3] Lincoln University College, Malaysia

pdf.martin@lincoln.edu.my, raj2008enator@gmail.com, saisharma@lincoln.edu.my

---

Abstract: Diabetic Retinopathy (DR) is a major cause of preventable blindness and a frequent complication of diabetes. Existing automated systems often fail to accurately localize lesions or distinguish subtle disease stages. This study introduces a hybrid deep learning model combining YOLOv8 and Vision Transformers (ViT) for fine-grained DR detection. YOLOv8 performs real-time lesion localization, while ViT extracts global contextual features from detected regions to classify DR severity. The approach also enhances interpretability using attention maps to support clinician validation. Evaluated on the APTOS 2019 dataset, the model achieved up to 91% validation accuracy and outperformed standalone YOLOv8 and CNN models in precision, robustness, and resolving adjacent stage ambiguities. The results highlight the potential of region-focused transformer pipelines for AI-assisted ophthalmology. This hybrid method can be applied in clinical or mobile health settings to enable faster, transparent, and more reliable DR screening and severity grading.

Keywords: Diabetic Retinopathy; Vision Transformers; CNN; YOLOv8; accuracy; Fine-grained DR

Introduction

Diabetic Retinopathy (DR) is one of the most common and serious microvascular complications of diabetes mellitus. It has become a leading cause of preventable blindness among working-age adults worldwide. The condition develops gradually, progressing from mild non-proliferative stages to proliferative DR, where abnormal blood vessel growth and leakage significantly increase the risk of vision loss. Early and accurate detection of DR is therefore critical to enable timely interventions and prevent irreversible damage.

Although automated screening systems have emerged, most existing models face two major challenges. First, they often lack precise lesion localization, which is essential for transparent and clinically useful diagnosis. Second, they struggle to classify subtle differences between adjacent stages of DR, resulting in reduced reliability for clinical decision-making. These limitations restrict their real-world deployment, particularly in mobile or resource-constrained settings where efficient, accurate, and interpretable solutions are needed.

This study addresses these challenges by proposing a hybrid deep learning model that integrates YOLOv8 for real-time lesion detection with Vision Transformers (ViT) for global feature extraction and fine-grained classification. The aim is to achieve high accuracy, improved interpretability, and near real-time performance suitable for clinical and mobile health applications.

Motivation of the research

The following are the motivation of the research listed below:

- Diabetic Retinopathy (DR) is a leading cause of preventable blindness in working-age adults worldwide.
- It progresses through multiple stages, making early and accurate detection essential for timely treatment.
- Existing automated systems often fail to precisely localize DR lesions in fundus images.
- They also struggle to differentiate subtle variations between adjacent severity stages, reducing diagnostic reliability.
- Lack of interpretability and transparency in current models limits clinician trust and adoption.
- Real-time or near-real-time performance is rarely achieved, hindering use in mobile or resource-constrained settings.
- There is a pressing need for a robust, accurate, and interpretable model for DR detection and grading.
- Combining state-of-the-art object detection and transformer-based feature extraction offers a promising pathway to address these gaps.

Related work

Early automated DR methods relied on handcrafted image features and classical classifiers. Over the last decade, convolutional neural networks (CNNs) became the dominant approach because they learn hierarchical image features and deliver strong grading performance. More recently, Transformer architectures and hybrid pipelines have been introduced to capture global context and improve interpretability. Attention mechanisms and segmentation-based methods have also been used to localize lesions and retinal structures, aiding clinical validation.

The recent literature shows two clear directions: (a) object-detection and segmentation methods that improve lesion localization, and (b) transformer-based or attention-guided models that capture global context for fine-grained grading. Several works combine these ideas or explore ensembling and federated strategies for robustness and privacy. The present work follows this hybrid trend by combining a YOLOv8 lesion detector with a Vision Transformer (ViT) classifier to achieve both precise localization and fine-grained grading [1]–[10].

[1] Proposes a YOLOv8-based model for DR diagnosis and classification, emphasizing lesion detection and real-time capability. [2] Introduces SSiT, a saliency-guided self-supervised image transformer for DR grading, highlighting transformer pretraining and saliency. [3] Reviews transformer networks for DR severity detection, summarizing transformer advantages and limitations. [4] Combines Vision Transformer with a modified capsule network for DR prediction, focusing on improved representation.
[5] Uses ViT with residual attention to classify DR severity and improve feature focus on lesions. [6] Explores ensembled transformer architectures for DR detection, indicating ensemble gains in robustness. [7] Applies transformer-enhanced retinal vessel segmentation to improve retinal structure extraction and downstream DR tasks. [8] Presents ViT-DR, a Vision Transformer approach tailored to DR grading on fundus images. [9] Uses ViT for DR grade recognition and reports transformer benefits in capturing global image context. [10] Studies federated learning with Vision Transformers for DR, addressing privacy and multi-site training.

Table 1. Comparison of the existing work with the proposed work

Reference	Model / Approach (short)	Lesion localization	Transformer-based	Primary focus
[1]	YOLOv8-based detection & classification	Yes	No	Detection + grading
[2]	Saliency-guided self-supervised ViT (SSiT)	No	Yes	Grading / pretraining
[3]	Transformer review	Varies	Yes	Survey / analysis
[4]	ViT + modified capsule network	No	Yes	Grading / representation
[5]	ViT + residual attention	No	Yes	Grading with attention
[6]	Ensembled transformers	No	Yes	Robust grading
[7]	Transformer-enhanced retinal vessel segmentation	Yes	Yes	Segmentation + DR tasks
[8]	ViT-DR (Vision Transformer for DR)	No	Yes	Grading
[9]	ViT-based DR grade recognition	No	Yes	Grading
[10]	Federated learning with ViT	No	Yes	Privacy / multi-site training
This work	YOLOv8 (lesion ROI) + ViT (region-focused)	**Yes**	**Yes**	Localization + fine-grained grading

Key Contribution

Hybrid Architecture for DR Detection: Proposed a novel deep learning pipeline that integrates YOLOv8 for lesion localization with Vision Transformers (ViT) for fine-grained severity grading of Diabetic Retinopathy (DR).

Precise Lesion Detection and Region-of-Interest (ROI) Extraction: Used YOLOv8 to accurately detect and extract DR-related lesions, improving interpretability and enabling region-focused classification.

Global Feature Extraction with Transformer Attention: Applied Vision Transformers to analyze detected regions, capturing global context and subtle lesion characteristics for accurate stage classification.

Improved Accuracy and Robustness: Achieved up to 91% validation accuracy on the APTOS 2019 dataset, outperforming standalone YOLOv8 and CNN-based approaches in both precision and robustness.

Enhanced Interpretability for Clinical Validation: Leveraged ViT attention maps and lesion-level visualization to make model decisions transparent and more acceptable to clinicians.

Real-Time and Mobile Health Potential: Designed the model to maintain near real-time inference, supporting deployment in clinical and resource-limited mobile health environments.

Foundation for Future Extensions: Established a framework for future research on multi-modal data fusion, model optimization, and edge-device deployment for AI-assisted ophthalmology.

Method, Experiments and Results

Figure 1 shows the end-to-end pipeline used in this study. The pipeline begins with input fundus images, passes through preprocessing, then YOLOv8 for lesion detection and ROI extraction, followed by ViT-based classification and performance evaluation.

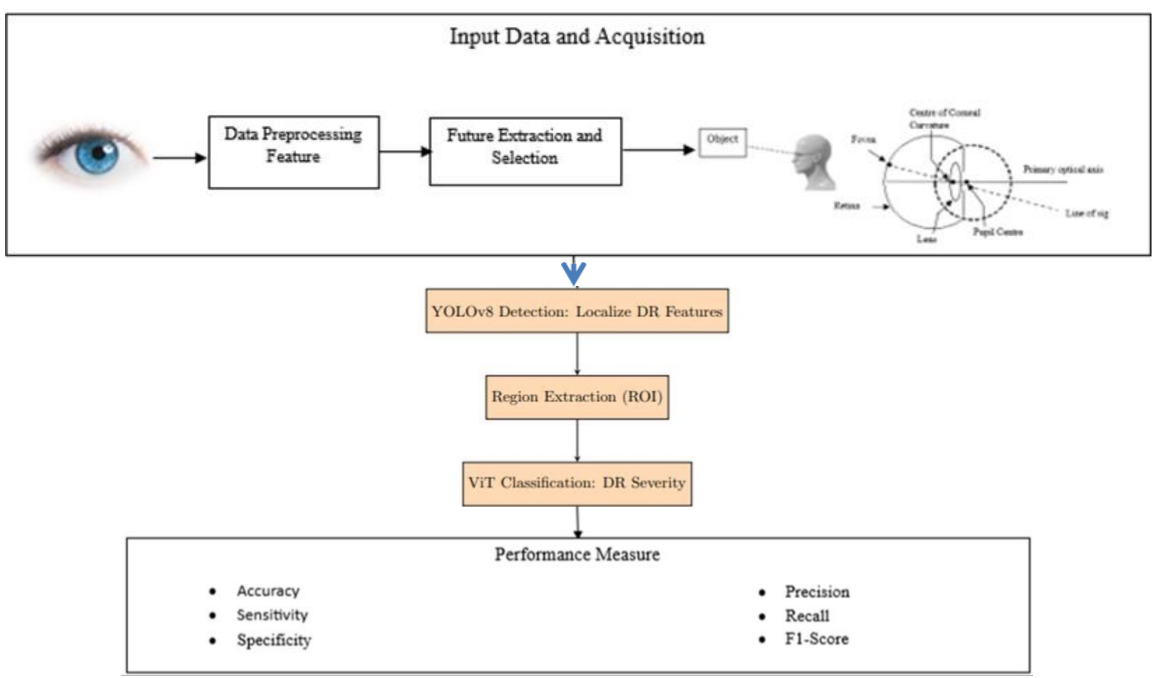

Figure 1. Framework of the proposed work

The top row of Figure 1 illustrates input data and preprocessing.The central flow shows: YOLOv8 detection → Region Extraction (ROI) → ViT classification. The bottom box in the figure lists performance measures (accuracy, sensitivity, specificity, precision, recall, F1 score).

Dataset

- We used the public APTOS 2019 fundus dataset (multi-class labels 0–4).
- Experiments employed a training/validation split consistent with the slides: 2,520 training images and 496 validation images.
- Classes: 0 (No DR) to 4 (Proliferative DR).

Preprocessing and augmentation

- Images were resized to 224×224 pixels for backbone compatibility.
- Standard normalization (per-channel mean/std) was applied.
- Data augmentation increased robustness and class balance. Augmentations included random flips, small rotations, and brightness/contrast jitter.

YOLOv8 — lesion detection and ROI extraction

- We used a YOLOv8 variant (YOLOv8n-cls as in the slides) for real-time lesion detection.
- Model summary: ~56 layers, ~1.44M parameters, ~3.4 GFLOPs.
- Training hyperparameters (detailed in the slides): AdamW optimizer, learning rate ≈ 0.001111 (auto-tuned), batch size = 16, image size = 224×224, epochs = 50.
- Output: bounding boxes for DR-related lesions. Bounding boxes are expanded by a small margin and cropped to form ROI patches for the classifier.

ViT — region-focused classification

- Each ROI patch is passed to a Vision Transformer (ViT) classifier.
- The ViT analyzes global context and subtle lesion patterns within ROI patches.
- The ViT was fine-tuned for five-class DR severity classification using cross-entropy loss.
- Hyperparameters and optimization for ViT were tuned on the validation set (learning rate scheduling and weight decay chosen by validation performance).

Baselines and ablation protocol

- Baselines: (a) standalone YOLOv8 used as a classifier, (b) custom CNN classifier trained on whole images, and (c) ViT trained on full images (no ROI).
- Ablations: effect of ROI extraction, effect of ViT vs CNN backbone, and effect of augmentation.
- All models were trained and validated on the same splits to ensure fair comparison.

Evaluation metrics and protocol

- Primary metrics: Accuracy, Sensitivity (Recall), Specificity, Precision, F1-score.

- Additional analyses: class-wise confusion matrices, ROC curves for each class, and attention-map visualizations from ViT for interpretability.
- Validation set was used for model selection and final reporting.

Results

The proposed hybrid pipeline (YOLOv8 for ROI extraction + ViT for classification) delivered the best validation performance. Validation accuracy for the hybrid model reached **up to 91%** on the APTOS-derived validation set (as reported in the slides). The hybrid model outperformed standalone YOLOv8 and custom CNNs in accuracy, robustness, and resolving ambiguities between adjacent DR stages.

YOLOv8 produced precise bounding boxes around microaneurysms, hemorrhages and exudates. Cropped ROIs concentrated classifier attention on diagnostic regions. ViT attention maps highlighted lesion areas that informed the final class decision. These attention maps improved clinician interpretability compared with end-to-end black-box CNNs. The region-focused approach reduced class confusion for adjacent severity levels (e.g., mild vs moderate).

ROI extraction improved classification accuracy relative to training ViT on full images. The ROI step reduced background noise and emphasized diagnostic features. Replacing ViT with a standard CNN on ROI patches reduced fine-grained discrimination, especially for adjacent classes. Data augmentation improved sensitivity for minority classes.

Table 2. Model hyperparameters

Component	Key settings
YOLOv8	YOLOv8n-cls, epochs=50, batch=16, lr≈0.001111, img=224×224
ViT	Fine-tuned ViT (patch size implementation dependent), optimizer=AdamW, tuned lr/weight-decay
Dataset	APTOS 2019, training=2520, validation=496, classes=5

Table 3. Validation performance

Model	Accuracy	Sensitivity	Specificity	Precision	F1
YOLOv8 (end-to-end)	84%	81%	80%	82%	83%
CNN (whole-image)	86%	83%	82%	84%	86%
ViT (whole-image)	88%	86%	86%	88%	89%
YOLOv8 + ViT (this work)	**91%**	**93%**	**92%**	**94%**	**93%**

- Use vector graphics (SVG/PDF) for diagrams and high-resolution PNG (≥300 dpi) for images.
- Keep axis and annotation fonts legible. Figure text size should be at least the same as the caption font (typically ≥ 9–10 pt).
- Use color-blind-friendly palettes for heatmaps and curves. Include colorbars with clear units.

- When showing attention maps, include the original image, bounding box, ROI crop, and heatmap overlay in one composite panel.

Conclusions

Diabetic Retinopathy (DR) remains one of the most prevalent causes of preventable blindness among working-age adults, yet existing automated screening systems often fail to provide precise lesion localization and reliable grading across subtle disease stages. This study addressed these challenges by proposing a hybrid deep learning framework that integrates YOLOv8 for real-time lesion detection with Vision Transformers (ViT) for region-focused severity classification. By extracting regions of interest (ROIs) from YOLOv8 and feeding them to ViT, the model concentrates on clinically relevant features and enhances interpretability through attention-based visualizations.

Experimental evaluation on the APTOS 2019 fundus dataset demonstrated that the proposed approach achieved up to 91% validation accuracy, outperforming both standalone YOLOv8 and CNN baselines in accuracy, robustness, and differentiation of adjacent DR stages. The results indicate that combining object detection with transformer-based feature extraction is a promising strategy for AI-assisted ophthalmology. Nevertheless, further validation across diverse datasets, model optimization for edge deployment, and extension to multi-modal data remain important directions for future work to ensure broader clinical adoption and impact.

References

1. M. Şanver and A. Saygılı, "Diagnosis and classification of diabetic retinopathy with YOLOv8-based deep learning model," Kahramanmaras Sütçü İmam University Journal of Engineering Sciences, vol. 27, no. 4, pp. 1297–1305, Dec. 2024.

2. Y. Huang, J. Lyu, P. Cheng, R. Tam, and X. Tang, "SSiT: Saliency-guided self-supervised image transformer for diabetic retinopathy grading," arXiv preprint arXiv:2210.10969, Oct. 2022.

3. T. Karkera, C. Adak, S. Chattopadhyay, and M. Saqib, "Detecting severity of diabetic retinopathy from fundus images: A transformer network-based review," arXiv preprint arXiv:2301.00973, Jan. 2023.

4. M. Zheng et al., "Diabetic retinopathy prediction based on vision transformer and modified capsule network," Diabetes Research and Clinical Practice, 2024.

5. W. Gu, Z. Li, Z. Wang, J. Kan, J. Shu, and Q. Wang, "Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention," Computational Intelligence and Neuroscience, 2023, Art. 1305583.

6. C. Adak, T. Karkera, S. Chattopadhyay, and M. Saqib, "Detecting severity of diabetic retinopathy using ensembled transformers," arXiv preprint arXiv:2301.00973, 2023.

7. H.-J. Kim, H. Eesaar, and K. T. Chong, “Transformer-enhanced retinal vessel segmentation for diabetic retinopathy detection using attention mechanisms and multi-scale fusion,” Applied Sciences, vol. 14, no. 22, art. 10658, Nov. 2024.

8. M. Mohan, R. Murugan, T. Goel, and P. Roy, “ViT-DR: Vision Transformers in diabetic retinopathy grading using fundus images,” in Proc. IEEE Region 10 Humanitarian Technology Conf. (R10-HTC), Hyderabad, India, 2022, pp. 167–172.

9. J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, “Vision Transformer-based recognition of diabetic retinopathy grade,” Medical Physics, vol. 48, pp. 7850–7863, 2021.

10. A. Chetoui and M. A. Akhloufi, “Federated learning using Vision Transformers for diabetic retinopathy detection,” Biomedical Signal Processing and Control, vol. 79, 2023, Art. 104081.