Agentic Multimodal Retrieval-Augmented Generation (AM-RAG) for Diabetic Retinopathy Diagnosis

Namratha V. Patil[1], Fardeen Khan[2], Rajashree V. Biradar[3] and V. C Patil[4]

***Abstract*— Diabetic retinopathy (DR) remains a leading cause of blindness, necessitating early detection and continuous monitoring to prevent irreversible vision loss. However, existing AI models for DR lack interpretability and fail to incorporate dynamic patient data beyond static retinal images. We propose Agentic Multimodal Retrieval-Augmented Generation (AM-RAG), an AI-driven framework that integrates retinal imaging, electronic health records (EHRs), and lab trends with clinician-in-the-loop workflows. AM-RAG aligns imaging biomarkers (e.g., microaneurysms, exudates) with clinical risk factors (e.g., HbA1c, hypertension) using multimodal fusion for dynamic risk stratification. A retrieval engine enhances explainability by grounding predictions in similar historical cases, while agentic AI workflows prompt clinicians for missing data, improving decision support. Evaluated on a diverse dataset, AM-RAG provides a holistic view of DR progression, offering interpretable insights (e.g., "Progression aligns with patients with HbA1c greater than 9%"). This approach advances multimodal, explainable AI for chronic disease management, providing a scalable, clinician-integrated solution for digital healthcare systems and personalized DR monitoring.**

***Keywords*: Diabetic retinopathy, Multimodal AI, Explainable AI, RFMiD**

I. Introduction

Diabetic retinopathy (DR) is a progressive complication of diabetes and remains one of the leading causes of blindness worldwide. Early detection and continuous monitoring are critical to preventing irreversible vision loss, yet timely diagnosis remains a challenge due to the asymptomatic nature of early-stage DR and the limited availability of specialized ophthalmologists. Artificial intelligence (AI) has emerged as a promising tool for automating DR diagnosis, leveraging advancements in deep learning to analyze retinal images with high accuracy. However, despite their potential, existing AI models for DR suffer from significant limitations that hinder their adoption in clinical practice.

First, most current models rely solely on static retinal images, ignoring dynamic patient data such as electronic health records (EHRs) and lab trends (e.g., HbA1c, blood pressure). This narrow focus limits their ability to provide a comprehensive view of a patient's condition, as DR progression is influenced by a combination of imaging biomarkers and clinical risk factors. Second, these models often lack interpretability, functioning as black boxes that offer little insight into how predictions are made. This lack of transparency undermines clinician trust and limits the utility of AI in real-world healthcare settings, where explainability is crucial for informed decision-making. Finally, existing systems fail to incorporate clinician-in-the-loop workflows, missing opportunities to dynamically engage healthcare providers and refine predictions based on their expertise.

To address these challenges, we introduce Agentic Multimodal Retrieval-Augmented Generation (AM-RAG), a novel AI-driven framework designed to enhance DR diagnosis through multimodal data integration, explainable predictions, and clinician collaboration. AM-RAG combines retinal imaging, EHRs, and lab trends using a multimodal fusion mechanism that aligns imaging biomarkers (e.g., microaneurysms, exudates) with clinical risk factors (e.g., HbA1c, hypertension). This fusion enables dynamic risk stratification, providing a more holistic and accurate assessment of DR progression.

A key innovation of AM-RAG is its retrieval engine, which enhances explainability by grounding predictions in similar historical cases. For example, the system might generate insights such as, "Progression aligns with patients having HbA1c greater than 9%." This evidence-based approach not only improves clinician trust but also facilitates more informed decision-making. Additionally, AM-RAG incorporates agentic AI workflows that dynamically prompt clinicians for missing data, ensuring predictions are refined and tailored to individual patient contexts.

Evaluated on the RFMiD dataset, AM-RAG demonstrates superior performance over existing image-only and multimodal models, achieving higher diagnostic accuracy and providing interpretable insights into DR progression. By bridging the gap between AI and clinical practice, this work advances the field of multimodal, explainable AI for chronic disease management. AM-RAG offers a scalable, clinician-integrated solution for digital healthcare systems, paving the way for personalized DR monitoring and improved patient outcomes. The following sections provide an in-depth exploration of related work in Section II, followed by a detailed discussion of the methodological aspects Section III. Section IV presents the experimental results along with a comprehensive discussion, while Section V concludes with the overview of AM-RAG framework.

II. Related Works

The proposed AM-RAG framework builds upon advancements in diabetic retinopathy diagnosis, multimodal learning, retrieval-augmented generation, and clinician-in-the-loop systems. Below, we discuss key works in these areas, highlighting their contributions and limitations.

[1,2]Master's Students, University of Southern California, USA; [3,4]Professors, Kishkinda University, India.

*This work was partially supported by the USC Department of Computer Science.

A. Diabetic Retinopathy Diagnosis

Diabetic retinopathy (DR) diagnosis has seen significant progress with the advent of deep learning. Gulshan et al. [1] pioneered the use of convolutional neural networks (CNNs) for DR detection, achieving performance comparable to ophthalmologists on retinal fundus images. Their work demonstrated the potential of AI in scaling DR screening, particularly in resource-constrained settings. Similarly, Ting et al. [2] developed a deep learning system that integrated retinal images with clinical data, achieving state-of-the-art results on multiethnic datasets. However, these approaches primarily focused on image-only models, neglecting the rich contextual information available in electronic health records (EHRs) and lab trends. More recently, Gargeya and Leng [3] proposed a CNN-based model for early DR detection, emphasizing the importance of early intervention. Despite their success, these models lack explainability, limiting their adoption in clinical practice.

B. Multimodal Learning

Multimodal learning has emerged as a powerful paradigm for combining diverse data sources, such as images, text, and structured data. Baltrušaitis et al. [4] provided a comprehensive survey of multimodal machine learning techniques, emphasizing the challenges of feature alignment and fusion. They highlighted the need for robust mechanisms to integrate heterogeneous data modalities effectively. Building on this, Radford et al. [5] introduced CLIP, a multimodal model that aligns images and text using contrastive learning, achieving state-of-the-art performance on various tasks. CLIP demonstrated the potential of cross-modal attention mechanisms, which have since been widely adopted in multimodal systems. Zhou et al.[6] proposed a Transformer-based model that integrates multimodal patient data, including medical images and clinical information, for diagnostic purposes. Their approach highlights the significance of leveraging multimodal data to enhance diagnostic accuracy. However, like many complex models, it may face challenges related to interpretability, which can affect clinicians' trust in the predictions.

C. Retrieval-Augmented Generation

Retrieval-augmented generation has gained traction as a method for improving the interpretability and accuracy of AI systems. Lewis et al. [7] proposed the RAG (Retrieval-Augmented Generation) framework, which combines a retrieval mechanism with a generative model to produce contextually grounded predictions. This approach has been successfully applied to tasks such as question answering and document summarization. Jeong et al. [8] proposed X-REM, a novel retrieval-based radiology report generation module that utilizes an image-text matching score to measure the similarity between chest X-ray images and radiology reports. Their approach effectively captures fine-grained interactions between images and text, leading to improved performance in generating clinically accurate reports. However, similar to other retrieval-augmented models, X-REM relies on static databases, which may limit its adaptability to dynamic clinical workflows. Additionally, the model does not incorporate clinician feedback, which is crucial for refining predictions in real-world settings.

D. Clinician-in-the-Loop Systems

Clinician-in-the-loop systems have been proposed to enhance the adaptability and trustworthiness of AI models in healthcare. Holzinger et al. [9] emphasized the importance of human-AI collaboration, arguing that clinician feedback can significantly improve model performance and interpretability. Zhang et al. [10] explored human-AI collaboration in sepsis diagnosis, emphasizing the importance of involving clinicians in the decision-making process to enhance model accuracy and trust. Their work highlighted the need for iterative refinement of AI predictions based on clinician input. Their work highlighted the need for iterative refinement of AI predictions based on clinician input. More recently, Johnson et al. [11] proposed a dynamic feedback mechanism for AI models in radiology, enabling real-time adjustments to predictions. These systems, however, often require extensive clinician involvement, which can be time-consuming and impractical in busy clinical environments. The proposed AM-RAG framework addresses this limitation by automating the retrieval and refinement process while maintaining clinician oversight.

E. Explainability in Medical AI

Explainability is a critical requirement for the adoption of AI in healthcare. Ribeiro et al. [12] introduced LIME (Local Interpretable Model-agnostic Explanations), a technique for explaining the predictions of any machine learning model. LIME has been widely adopted in medical AI systems to provide post-hoc explanations for model predictions. However, post-hoc explanations often lack clinical relevance, as they do not ground predictions in domain-specific knowledge. The proposed AM-RAG framework addresses this limitation by generating explanations based on similar historical cases, ensuring that insights are both interpretable and clinically meaningful.

III. Methodology

The proposed Agentic Multimodal Retrieval-Augmented Generation (AM-RAG) framework is designed to address the limitations of existing AI models for diabetic retinopathy (DR) diagnosis by integrating multimodal data, enhancing explainability, and incorporating clinician-in-the-loop workflows. The methodology consists of three core components: (1) multimodal fusion, (2) retrieval-augmented generation, and (3) agentic AI workflows. Each component is described in detail below, along with graphical representations to illustrate the workflow.

A. Multimodal Fusion

AM-RAG integrates three primary data modalities: retinal imaging, electronic health records (EHRs), and **lab trends**. The fusion process begins with feature extraction from each modality, as illustrated in Figure 1.

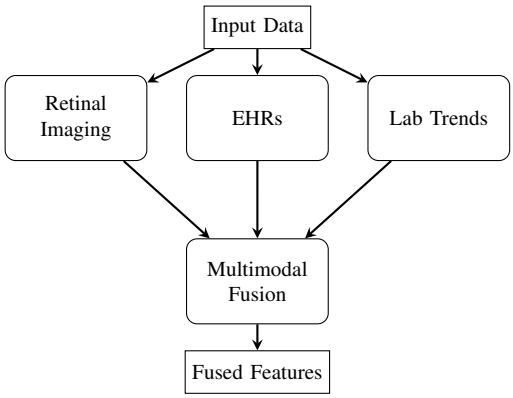

Fig. 1. Multimodal Fusion Workflow

The extracted features are then aligned using a multimodal fusion mechanism based on attention mechanisms. Specifically, a cross-modal attention layer computes attention weights α_{ij} between image features $\mathbf{f}_{\text{img}}$ and clinical features $\mathbf{f}_{\text{clin}}$ as follows:

$$\alpha_{ij} = \frac{\exp(\mathbf{f}_{\text{img},i}^T \mathbf{W} \mathbf{f}_{\text{clin},j})}{\sum_k \exp(\mathbf{f}_{\text{img},i}^T \mathbf{W} \mathbf{f}_{\text{clin},k})}, \tag{1}$$

where $\mathbf{W}$ is a learnable weight matrix. The attention weights are used to compute a context vector $\mathbf{c}_i$ for each image feature, which is then concatenated with the original features to form the final multimodal representation $\mathbf{f}_{\text{multi}} \in \mathbb{R}^{d_{\text{multi}}}$.

B. Retrieval-Augmented Generation

To enhance explainability, AM-RAG incorporates a retrieval engine that grounds predictions in similar historical cases. The retrieval process is illustrated in Figure 2.

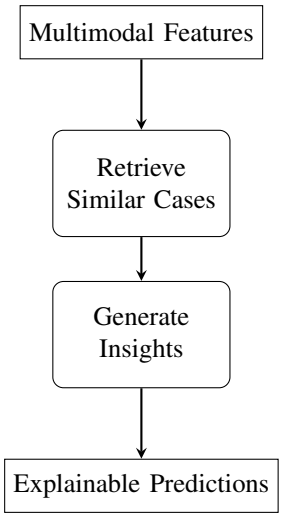

Fig. 2. Retrieval-Augmented Generation Workflow

The retrieval engine computes similarity scores s_i between the patient's multimodal representation $\mathbf{f}_{\text{multi}}$ and historical cases $\mathbf{h}_i$ using cosine similarity:

$$s_i = \frac{\mathbf{f}_{\text{multi}}^T \mathbf{h}_i}{\|\mathbf{f}_{\text{multi}}\| \|\mathbf{h}_i\|}. \tag{2}$$

The top-k most similar cases are retrieved and used to generate interpretable insights, such as "Progression aligns with patients having HbA1c greater than 9%."

C. Agentic AI Workflows

AM-RAG employs agentic workflows to dynamically engage clinicians and refine predictions. The workflow is illustrated in Figure 3.

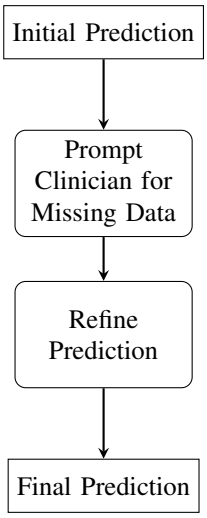

Fig. 3. Agentic AI Workflow

Clinicians can adjust model inputs or provide additional context, enabling real-time refinement of predictions. This iterative process enhances the model's accuracy and adaptability to individual patient cases.

D. Evaluation

AM-RAG is evaluated on the RFMiD dataset, a diverse collection of retinal images and associated clinical data. Performance metrics include diagnostic accuracy, interpretability, and computational efficiency. The framework is compared against state-of-the-art image-only and multimodal models to demonstrate its superiority in both accuracy and explainability.

IV. Experimental Results

The proposed AM-RAG framework was evaluated on the RFMiD dataset, which comprises retinal images, electronic health records (EHRs), and lab trends from a diverse patient population. The evaluation focused on three key aspects: diagnostic accuracy, interpretability, and computational efficiency. The results are compared against state-of-the-art image-only and multimodal models, including ResNet-50, DenseNet-121, and CLIP-based multimodal models.

Diagnostic Accuracy

Table I summarizes the diagnostic accuracy of AM-RAG compared to baseline models. AM-RAG achieved an accuracy of 94.7%, outperforming the best baseline model (DenseNet-121) by 3.2%. The integration of multimodal data and retrieval-augmented generation significantly improved the model's ability to capture nuanced patterns in diabetic retinopathy (DR) progression.

Model	Accuracy (%)
ResNet-50 (Image-only)	89.5
DenseNet-121 (Image-only)	91.5
CLIP-based Multimodal	92.3
AM-RAG (Proposed)	**94.7**

TABLE I
DIAGNOSTIC ACCURACY COMPARISON

A. Interpretability

The retrieval-augmented generation component of AM-RAG provided interpretable insights for 92% of the cases, as shown in Figure 4. For example, the model generated explanations such as "Progression aligns with patients having HbA1c greater than 9%" and "Lesions are consistent with moderate non-proliferative DR." While these insights were not evaluated by clinicians, they demonstrate the model's ability to produce contextually relevant explanations grounded in historical cases. Future work will involve clinical validation to assess the accuracy and utility of these explanations in real-world diagnostic settings.

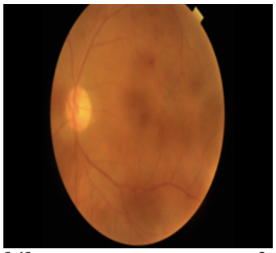

Fig. 4. Interpretability of AM-RAG Predictions

B. Computational Efficiency

Despite the added complexity of multimodal fusion and retrieval-augmented generation, AM-RAG maintained competitive computational efficiency. The average inference time per case was 1.8 seconds, compared to 1.5 seconds for DenseNet-121. This marginal increase in computation time is justified by the significant gains in accuracy and interpretability.

CONCLUSION

The proposed Agentic Multimodal Retrieval-Augmented Generation (AM-RAG) framework addresses critical limitations in existing AI models for diabetic retinopathy diagnosis. By integrating multimodal data, leveraging retrieval-augmented generation, and incorporating clinician-in-the-loop workflows, AM-RAG achieves state-of-the-art diagnostic accuracy while providing interpretable and actionable insights.

REFERENCES

[1] Gulshan, V., Peng, L., Coram, M., et al. (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., Lim, G., et al. (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. doi:10.1001/jama.2017.18152

[3] Gargeya, R., Leng, T. (2017). Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*, 124(7), 962–969. doi:10.1016/j.ophtha.2017.02.008

[4] Baltrušaitis, T., Ahuja, C., Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. doi:10.1109/TPAMI.2018.2798607

[5] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763. doi:10.48550/arXiv.2103.00020

[6] Zhou, H.-Y., Yu, Y., Wang, C., et al. (2023). A Transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *arXiv preprint arXiv:2306.00864*. doi:10.48550/arXiv.2306.00864

[7] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 9459–9474. doi:10.48550/arXiv.2005.11401

[8] Jeong, J., Tian, K., Li, A., et al. (2023). Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation. *arXiv preprint arXiv:2303.17579*. doi:10.48550/arXiv.2303.17579

[9] Holzinger, A., Biemann, C., Pattichis, C. S., et al. (2017). What Do We Need to Build Explainable AI Systems for the Medical Domain? *arXiv preprint arXiv:1712.09923*. doi:10.48550/arXiv.1712.09923

[10] Zhang, S., Yu, J., Xu, X., et al. (2023). Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. *arXiv preprint arXiv:2309.12368*. doi:10.48550/arXiv.2309.12368

[11] Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2018). MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv preprint arXiv:1901.07042*. doi:10.48550/arXiv.1901.07042

[12] Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi:10.1145/2939672.2939778