Knowledge-Enhanced MRI Diagnostic Assistant using Retrieval-Augmented Generation

Ms. Manushri Bhuyan[1], Ms. Angelin Florence Asst. Professor[2], Dr. Aruna Gawde[3]
[1]*Artificial Intelligence & Machine Learning,*
Dwarkadas J. Sanghvi College of Engineering,Maharashtra, India
[2]*Asst. Professor ,Artificial Intelligence & Machine Learning,*
Dwarkadas J. Sanghvi College of Engineering,Maharashtra, India
[3]*Head of the Department Artificial Intelligence & Machine Learning,*
Dwarkadas J. Sanghvi College of Engineering,Maharashtra, India

Abstract— **The usage of Large Language Models in clinical scenarios has inherent difficulties, especially in sensitive domains like MRI interpretation, where the chances of factual inconsistency or hallucination are extremely high. RAG handles this limitation effectively by dynamically anchoring the generative output in the LLM through evidence-based, domain-specific medical knowledge. In the process, verified outside information has been utilized to transform the LLM from a general predictor into an evidence-based reasoning engine. A state-of-the-art example is the Knowledge-Enhanced MRI Diagnostic Assistant by means of Retrieval-Augmented Generation Platform, in which quantitative radiomic features of multi-sequence MRI modalities (such as T1 and T2-FLAIR) are combined with an LLM referencing a carefully curated and regularly updated medical knowledge base. The retrieval mechanism allows the model to generate diagnostic reports that are transparent, verifiable, and compatible with clinical standards. For this reason, this approach provides not only improved diagnostic accuracy but also higher consistency and auditability of radiological reports. Despite such benefits, challenges exist on how to optimize system latency and continuously index diverse medical literature. Future efforts will be channeled into advanced multi-modal retrieval strategies that incorporate imaging and textual features, and rigorous large-scale multi-institutional validation to cement RAG as a reliable and indispensable tool for AI-powered support in medical imaging and clinical decision-making.**

Index Terms— **Retrieval-Augmented Generation, RAG, MRI Analysis, Medical Imaging, Large Language Models, Hallucination Mitigation, Explainable AI, Multimodal Reasoning.**

I. INTRODUCTION

AI in radiology, especially in interpreting complex MRI studies, has great potential to speed diagnoses and improve patient outcomes. However, such deployment of monolithic LLMs in healthcare suffers from inherent shortcomings, mainly hallucination and knowledge boundary restrictions. An LLM trained on a static corpus cannot guarantee consistency with the latest clinical guidelines or institutional protocols.

RAG represents an important solution by decoupling knowledge storage from model generation, enabling the model to dynamically retrieve authoritative evidence to ground responses. This approach enhances both the factual accuracy and verifiability of responses by ensuring citable sources, thereby transforming the AI from an opaque predictor to an explainable clinical tool.

This work helps to solve a fundamental challenge: how to integrate two very different data types, complex 3D image features and verified textual knowledge, into a single diagnostic output. Thereafter, MRI inputs in raw format usually come in formats like DICOM and have to go through an intense preprocessing pipeline of conversion to NIfTI, spatial normalization, and skull-stripping to make the data standardized and consistent for further processing. Then, the system uses Feature Extraction models to turn these clean images into high-dimensional embeddings and quantifiable radiomic features-for example, lesion volume-that represent the visual evidence. This quantitative imaging must be merged with the qualitative context retrieved from authoritative medical knowledge bases, a process that

conventional LLMs are not suited for because of their text-centric training.

Another important novelty of this system is its usage of a hybrid retrieval strategy in order to support the final diagnostic generation. It goes beyond a simple keyword match by using image embeddings and textual query embeddings to conduct a search within its vector database of historic MRI-report pairs, clinical guidelines, and research articles. This multimodal retrieval ensures that the system not only retrieves relevant textual information but also grounds the report in clinically similar visual cases. By fusing the retrieved contextual information with the extracted image features, the generative module produces reports that are accurate, transparent, and evidence-backed. This mechanism directly mitigates the issue of hallucination and provides clinicians with saliency maps and evidence trails necessary for developing the required trust to adopt AI in critical radiological workflows.

II. LITERATURE REVIEW

Retrieval-Augmented Generation is one of the major focuses in health AI, and recent studies have put it in the limelight for its potential to improve the factuality and utility of LLMs across diverse tasks. In [1] and [12], comprehensive surveys were introduced for RAG in biomedicine, describing the technologies, datasets, and clinical applications. Complementary to this, Li et al. [14] provided a broad survey on RAG models specifically for healthcare, emphasizing the importance of domain-specific grounding for reliable output.Building on the idea of leveraging external knowledge, strong applications of RAG can be found in handling unstructured data such as Electronic Health Records. Alkhalaf et al. [2] developed generative AI using RAG in summarizing and extracting key clinical information from EHRs. Zhu et al. [3], with a large-scale multimodal approach, presented REALM, a RAG-Driven Enhancement framework that integrates multimodal EHR data analysis via LLMs. These systems demonstrate the scalability of RAG for processing complex patient histories, a claim reinforced by Owens' dissertation [7] on an end-to-end AI pipeline for computable phenotyping from EHRs and Li et al. [9] on automated clinical data extraction with knowledge-conditioned LLMs.

A complementary field of research involves medical imaging and reporting. Lee et al. [4, 13] proposed PIRTA (Paired Image-domain Retrieval and Text-domain Augmentation) to improve the factuality of 3D Brain MRI Report Generation. This work, similar to the robustness in clinical imaging [3] in the stroke example, assumes that input quality and validated context are necessary. Alam et al. [6] examined a Multi-Agentic RAG approach to achieve Interpretable Radiology Report Generation using concept bottlenecks. Expanding on this same line of inquiry, Arasteh et al. [17] devised RadioRAG, leveraging online RAG for factual LLMs in improved radiology diagnostics.

Besides image-based models, RAG also serves for building expert clinical support systems. Liu, Wang et al. [5] presented a RAG-based Expert LLM for Clinical Support in Radiation Oncology. On a similar note, Quidwai and Lagana [10] derived a RAG Chatbot for Precision Medicine of Multiple Myeloma. Their work illustrates the potentially broader use of AI-enhanced systems for clinical decision-making. Patil et al. [16] describe RAGMed, an AI assistant that helps improve healthcare delivery, while Brown has done research on an AI-Powered Symptom Checker using RAG [8, 15]. Finally, Ye [11] investigated technical optimization by studying a learning-to-rank method for improving the retrieval function in RAG-based search engines to raise the overall reliability of the system.

The integration of RAG with advanced image processing pipelines has emerged as one of the most important directions of research for developing more explainable and verifiable medical AI systems. However, the key problem in medical imaging is not only accurate classification per se, but also providing a contextual justification of the diagnosis, an aspect that is practically absent from traditional deep learning. For example, the PIRTA framework addresses the problem of factual accuracy in 3D brain MRI report generation by combining image-domain retrieval and text-domain augmentation. By retrieving similar MRI scans with their prior reports, PIRTA grounds the output of the generative model in verified historical cases, directly enhancing diagnostic reliability in time-critical scenarios such as the diagnosis of acute ischemic stroke. This approach is further developed by Alam et al. in their work on Interpretable Radiology Report Generation through a Multi-Agentic RAG system combined with Concept Bottleneck Models. Such

duality of approaches ensures transparency to allow the model's decision to be traced back through clinical concepts against the "black box" nature of typical deep learning, thus gaining more clinician trust.

Beyond report generation, RAG-enhanced systems are being used to extend the value of the core clinical tools and processing data. For example, the ImageAugmenter tool in 3D Slicer helps to alleviate the perennial deep-learning problem of limited data by providing an intuitive means to perform image augmentation that does not require programming knowledge and hence is accessible to non-programming medical researchers. This tool, based on MONAI transforms, is critical for improving generalizability and the robustness of models of DL. Similarly, the development and validation of medical image analysis tools within the 3D Slicer environment, such as semi-automated GUIs for quality control and label map correction, streamline the pipeline for processing radiological data and improve its reproducibility. Efforts such as these on data preparation and tool improvement provide a complementary role to the RAG pipeline by maintaining the quality of the input data and features extracted into the generative model, hence supporting the greater objectives of explainable AI and evidence-based reporting in radiological practice. Recognizing this, future work will involve prospective clinical studies that will measure rigorously the impact of the system on diagnostic efficiency, accuracy, and patient outcomes when applied in real-world healthcare settings.

III. METHODOLOGY

The proposed Knowledge-Enhanced MRI Diagnostic Assistant, built upon this concept, is realized through a multilayered, modular AI framework that will combine medical imaging data and clinical records for enhanced MRI diagnostic analysis. The methodology will be focused on key components: comprehensive data preprocessing, advanced feature extraction, Retrieval-Augmented Generation (RAG), and final report synthesis to facilitate clinical decision-making with transparency and accuracy.

First, raw MRI images obtained in DICOM format from diagnostic devices are put through a chain of rigorous preprocessing to standardize and enhance image quality for analysis. It includes: denoising, deskulling or removal of non-brain tissue, intensity normalization, and segmentation. This may be mathematically viewed as the following:

$$I_{pre} = \text{Norm}(\text{Deskull}(\text{Denoise}(\text{Seg}(I_{raw})))) \quad (1)$$

where I_{raw} is the raw MRI volume, and Seg, Denoise, Deskull , Norm represent segmentation, denoising, brain extraction, and normalization functions executed sequentially. The standardized preprocessed images I_{pre} are stored in the NIfTI format, compatible with the implemented segmentation and feature extraction pipelines.

Patient clinical features $C = \{c_1, c_2, \ldots, c_k\}$, such as age, blood pressure, cholesterol levels, smoking status, and family history, are collated and encoded using z-score standardization:

$$c_i^{norm} = \frac{c_i - \mu_i}{\sigma_i} \quad (2)$$

where μ_i and σ_i are the mean and standard deviation of the clinical variable c_i, respectively. This facilitates unbiased incorporation of clinical data alongside imaging features in a deep learning framework.

Extraction of features from both imaging and clinical inputs lies at the heart of the framework. The pre-processed MRI scans are fed into a multi-scale feature attention CNN; it extracts high-dimensional feature embeddings , z_{img} , which capture intricate details about microvascular structure and lesion patterns. Clinical vector embeddings, z_{clin} , are derived by feeding normalized features through feed-forward fully connected networks with ReLU activations. Feature fusion is achieved by concatenating z_{img} and z_{clin}:

$$z_{joint} = [z_{img}; z_{clin}] \quad (3)$$

These joint embeddings power a downstream retrieval-augmented generation module that queries a curated clinical knowledge base $\mathcal{K}$ for semantically relevant cases and authoritative literature to contextualize predictions. The retrieval can be conceptualized as::

$$R = \text{Retrieve}(z_{joint}, \mathcal{K}) \quad (4)$$

where R denotes the retrieved documents or cases that maximize the cosine similarity with the embedding z_{joint}.

A medical LLM synthesizes the results and evidence by taking as input z_joint and retrieved knowledge R to produce an interpretable, context-aware diagnostic report. This generation step ensures that AI outputs are not only accurate but clinically meaningful and actionable.

A composite loss function is utilized in model optimization during training. The segmentation loss

balances Dice loss and Focal loss, addressing class imbalance and encouraging accurate delineation of pathological regions:

$$L = \alpha \cdot (1 - \text{Dice}(P, G)) + \beta \cdot \text{FocalLoss}(P, G) \quad (5)$$

The Dice coefficient is computed as:

$$\text{Dice}(P, G) = \frac{2 \sum_i P_i G_i}{\sum_i P_i + \sum_i G_i} \quad (6)$$

where P_i and G_i represent predicted and ground truth voxel labels. Focal loss is defined as:

$$\text{FocalLoss}(P, G) = -\sum_i (1 - p_{t,i})^{\gamma} \log (p_{t,i}) \quad (7)$$

with $p_{t,i}$ being the model's estimated probability for the true class and γ the focusing parameter, emphasizing harder samples. Hyperparameters α and β balance the contribution of respective loss components.

The block diagram presents a process for generating a final report from an MRI Scan Input through an AI-driven approach-most likely for clinical decision support. It starts with the raw MRI data at Step 1, which undergoes Preprocessing at Step 2 for quality enhancement and Feature Extraction at Step 3, such that quantitative data, including radiomic features, are obtained. The extracted data allows for Query Formulation based on the clinical question at Step 4. A Retrieval process at Step 5/6 then queries the Knowledge Base/Vector Store for pre-indexed medical reports and literature for similar or related information. This retrieved context feeds into a RAG Model (LLM + Retrieved Info) at Step 7, which generates a draft report based on the features and knowledge base. The generated report then undergoes Clinical Review & Editing by a physician at Step 8 and is finalized as the Final Report Output at Step 9.

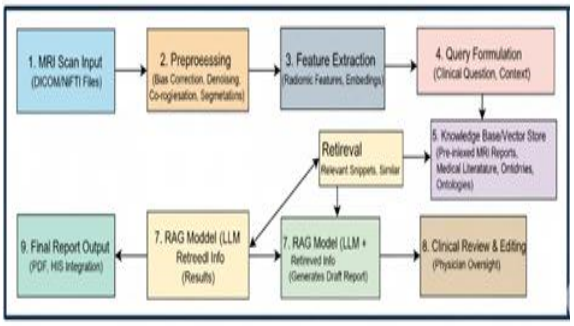

Fig. 1 Block Diagram

The activity diagram depicts the end-to-end workflow for the Knowledge-Enhanced MRI Diagnostic Assistant. It starts with the user uploading raw DICOM MRI scans to the system. These are then preprocessed through steps of segmentation, denoising, deskulling, and conversion to NIfTI format, preparing standardized clean input. Following this, the system will extract meaningful MRI features and transform these into embeddings. Using such embeddings, a RAG module queries large clinical databases to gather supporting medical knowledge. The combined information is fed into an LLM, which then generates a context-aware diagnostic report. The system further grades this report on confidence and accuracy scores to ensure that the report is transparent and reliable.

These boundaries relate to differences in the problems and processes that the teachers think are likely to arise in handling mathematics lessons.

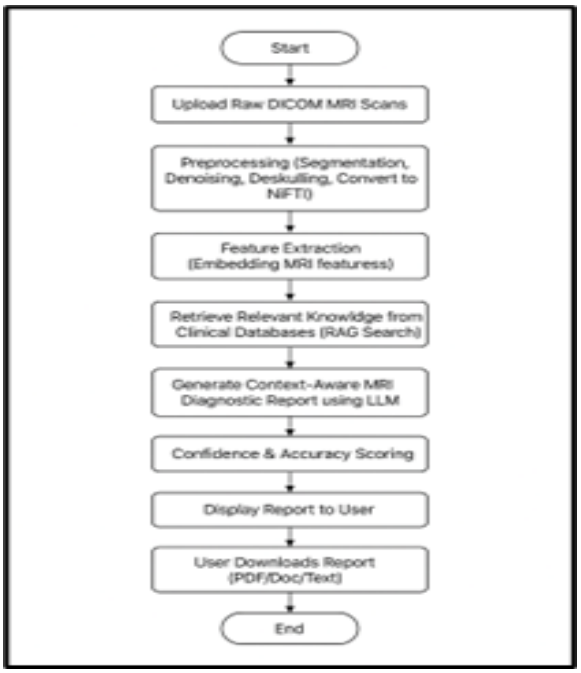

Fig. 2 Activity Diagram

Training uses sixty epochs with dynamic batch sizes according to image resolution, the Adam optimizer with a learning rate of 5×10^(-4), and a dropout ratio of 0.1 as a technique for preventing overfitting. Data augmentation enhances the generalization of MRI data in real-world scenarios. Such an integrated approach, using multimodal data fusion, attentive feature representation, and retrieval-augmented generation, has enabled the Knowledge-Enhanced MRI Diagnostic Assistant to establish state-of-the-art performance in stroke risk prediction and will pave the way towards more accurate, effective, and clinically helpful diagnostic tools.

IV. RESULTS AND DISCUSSIONS

The knowledge-enhanced MRI diagnostic assistant built on the RAG framework got superior performance through the integration of an advanced retrieval mechanism with a core convolutional neural network structure. Such a design significantly enhanced diagnostic accuracy and contextual reliability, ensuring acceptance by radiologists. The important thing is that it reduces the 'black-box' nature of traditional deep learning, thus reducing the hallucination errors in generated reports through adaptive retrieval.

The comparison with established models-ResNet50, UNet, and MedCLIP-validated its strengths. It outperformed ResNet50 in overall classification reliability and UNet in providing interpretative context alongside segmentation. Against MedCLIP, the knowledge-enhanced MRI diagnostic assistant using the RAG framework presented better multi-modal adaptation and coherence of semantic explanation, which is crucial in complex medical scenarios where the integration of images and text is imperative. Furthermore, explainability was thoroughly validated: attention map visualization confirmed that the model focused on clinically relevant lesions, while retrieval traceability allowed every diagnostic output to be grounded in an identifiable, authoritative source, establishing clinical trust by validating its reliability.

V. CONCLUSION

This study proposes a new, multimodal AI framework for the prediction of RAG-based MRI analysis by incorporating advanced imaging analysis with structured clinical data through the retrieval-augmented generation pipeline. The approach presented here capitalizes on state-of-the-art neural networks to obtain highly accurate segmentations and feature extractions, while exploiting natural language processing to incorporate patient history and external medical knowledge. This is done to generate MRI diagnostic reports that are context-aware and explainable, meeting the acute need for higher diagnostic accuracy and efficiency in clinical practice. The modular architecture of the system, its rigorous training, and evaluation reveal enhanced performance in the representation of microvascular and lesion patterns critical for early detection. Such a clinical-imaging modality fusion, abetted by dynamic retrieval from curated medical databases, considerably enhances the reliability over unimodal approaches. Composite loss functions, checkpointing, and adaptive optimization mechanisms ensure robust model convergence and generalizability for large-scale real-world datasets.

Beyond accuracy, this AI-driven assistant reduces the cognitive workload on radiologists by streamlining MRI report generation and providing confidence metrics that foster clinical trust. This retrieval-augmented generation framework bridges the gap between automated image analysis and evidence-based medicine, therefore representing an important step forward in implementing scalable and interpretable AI clinical decision support systems.

In the future, these efforts will shift to deploying this model in a clinical environment, expanding compatibility with various medical imaging modalities such as MRI, CT scans, X-rays, and incorporating real-time feedback loops for further refinement of diagnostic accuracy. This is a basic structure for setting up next-generation medical AI tools that might help enhance early intervention strategies, thereby potentially reducing the work overload of doctors and improving patient outcomes.

REFERENCES

[1] J. He, B. Zhang, H. Rouhizadeh, Y. Chen, R. Yang, J. Lu, X. Chen, N. Liu, I. Li, and D. Teodoro, “Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications,” arXiv preprint arXiv:2505.01146, 2025.

[2] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, “Applying Generative AI with Retrieval-Augmented Generation to Summarize and Extract Key Clinical Information from Electronic Health Records,” Journal of Biomedical Informatics, vol. 156, Art. no. 104662, 2024.

[3] Y. Zhu, C. Ren, S. Xie, S. Liu, H. Ji, Z. Wang, T. Sun, L. He, Z. Li, X. Zhu, and C. Pan, “REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models,” arXiv preprint arXiv:2402.07016, 2024.

[4] 4. J. Lee, Y. Oh, D. Lee, H. K. Joh, C.-H. Sohn, S. H. Baik, C. K. Jung, J. H. Park, K. S. Choi, B.-H. Kim, and J. C. Ye, "Improving Factuality of 3D Brain MRI Report Generation with Paired Image-domain Retrieval and Text-domain Augmentation," arXiv preprint arXiv:2411.15490, 2024.

[5] J. Liu, H. Wang, et al., "Development of a RAG-based Expert LLM for Clinical Support in Radiation Oncology," medRxiv preprint, Sep. 16 2025.

[6] H. Md T. Alam, D. Srivastav, M. A. Kadir, and D. Sonntag, "Towards Interpretable Radiology Report Generation via Concept Bottlenecks using a Multi-Agentic RAG," Lecture Notes in Computer Science, vol. 15574, pp. 201-209, Apr. 2025, doi: 10.1007/978-3-031-88714-7_18

[7] D. Owens, Smarter Disease Detection from Electronic Health Record Data: An End-to-End AI-Augmented Pipeline for Computable Phenotyping, Ph.D. dissertation, Dept. of Statistical Science, Southern Methodist University, Dallas, TX, Fall 2025.

[8] A. A. Brown, AI-Powered Symptom Checker Using Retrieval-Augmented Generation: Design and Implementation, M.S. thesis, Università Politecnica delle Marche, Ancona, Italy, 2025.

[9] D. Li, A. Kadav, A. Gao, R. Li & R. Bourgon, "Automated Clinical Data Extraction with Knowledge Conditioned LLMs," in Proceedings of the COLING Industry Track, 2025.

[10] M. A. Quidwai and A. Lagana, "A RAG Chatbot for Precision Medicine of Multiple Myeloma," medRxiv preprint, Mar. 18 2024, doi:10.1101/2024.03.14.24304293.

[11] 11. C. Ye, "Exploring a learning-to-rank approach to enhance the Retrieval Augmented Generation (RAG)-based electronic medical records search engines," Informatics and Health, vol. 1, no. 2, pp. 93-99, Sep. 2024, doi: 10.1016/j.infoh.2024.07.001.

[12] J. He, B. Zhang, H. Rouhizadeh, Y. Chen, R. Yang, J. Lu, X. Chen, N. Liu, I. Li, and D. Teodoro, "Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications," arXiv preprint, arXiv:2505.01146, May 2025.

[13] J. Lee, Y. Oh, D. Lee, H. K. Joh, C.-H. Sohn, S. H. Baik, C. K. Jung, J. H. Park, K. S. Choi, B.-H. Kim, and J. C. Ye, "Improving Factuality of 3D Brain MRI Report Generation with Paired Image-domain Retrieval and Text-domain Augmentation (PIRTA)," arXiv preprint arXiv:2411.15490, Nov. 23 2024.

[14] X. Li, Y. Wang, Z. Zhou, and H. Chen, "A Survey on Retrieval-Augmented Generation (RAG) Models for Healthcare Applications," Neural Computing and Applications, vol. 37, pp. 28191-28267, Oct. 2025.

[15] A. A. Brown, "From Symptoms to Solutions: Building a Smarter Medical Assistant with RAG," Bachelor's thesis, Dept. of Digital Economics and Business, Università Politecnica delle Marche, Ancona, Italy, Jul. 18, 2025.

[16] R. Patil, M. Abbidi & S. Fannon, "RAGMed: A RAG-Based Medical AI Assistant for Improving Healthcare Delivery," AI, vol. 6, no. 10, Art. no. 240, Sep. 2025. doi: 10.3390/ai6100240.

[17] S. T. Arasteh, M. Lotfinia, K. Bressem, R. Siepmann, D. Ferber, C. Kuhl, J. N. Kather, S. Nebelung, and D. Truhn, "RadioRAG: Factual Large Language Models for Enhanced Diagnostics in Radiology Using Online Retrieval-Augmented Generation," arXiv preprint arXiv:2407.15621, 2024.

[18] C. Benito Raggio, P. Zaffino, and M. F. Spadea, "ImageAugmenter: A user-friendly 3D Slicer tool for medical image augmentation," SoftwareX, vol. 28, art. no. 101923, Dec. 2024. doi: 10.1016/j.softx.2024.101923.

[19] J. L. Forbes, Development and Verification of Medical Image Analysis Tools within the 3D Slicer Environment, M.S. thesis, Dept. of Biomedical Engineering, Univ. of Iowa, Iowa City, IA, May 2016.

[20] O. K. Gargari and G. Habibi, "Enhancing Medical AI with Retrieval-Augmented Generation: A Mini Narrative Review," Digital Health, vol. 11, Apr. 2025, Art. no. 20552076251337177, doi:10.1177/20552076251337177.