From Data RAGs to Riches: Harmonizing MRI Sequence Classification with LLMs

Briana Malik, Researcher, St. Jude Children's Research Hospital
Dharmam Savani, MS; Asim Bag, MBBS, MD, EDiNR; Silu Zhang, PhD; Paul Yi, MD

Introduction

MRI sequence classification in pediatric neuroimaging is hindered by inconsistent DICOM header metadata, complicating multi-institution clinical trials. To address this, we developed a Retrieval-Augmented Generation (RAG) framework using LLMs and evaluated if such a framework could harmonize heterogeneous metadata across diverse imaging protocols and scanners – offering a data-agnostic, scalable solution.

Hypothesis

RAG-based-LLM will improve pediatric MRI sequence classification.

Methods

Our end-to-end RAG pipeline facilitates pediatric MRI sequence classification by retrieving relevant context from a vector database before using GPT-4o for decision-making. As DICOM metadata is often inconsistent/incomplete, providing additional context using RAG theoretically helps resolve ambiguities.
The pipeline included four stages (Figure 1):

1. Extracting DICOM headers
2. Tokenizing metadata into embeddings
3. Storing embeddings in a FAISS vector database
4. Querying with and without RAG to compare classification performance

The dataset comprised 357 pediatric brain MRI series spanning T1, T2 (pre-/post-contrast), FLAIR, ADC, and “other” sequences, collected across heterogeneous scanners/protocols. Experiments were conducted using all 27 DICOM tags and the SeriesDescription tag alone to evaluate robustness. Accuracy of our RAG-enhanced pipeline was compared to baseline models using 1) all DICOM tags without RAG and 2) Series Description tags alone (with and without RAG).

Results

With all 27 tags, our RAG-based model achieved 83.7% accuracy, outperforming the non-RAG baseline (72.7%). In contrast, using only Series Description yielded lower accuracy (58.8% with RAG vs. 57.7% without). Most misclassifications occurred in the “other” category, often overlapping with ADC characteristics, reflecting the complexity of subtle sequence variations. Nonetheless, RAG reduced errors across all categories, including 17% improvement in T1 post-contrast classification.

Conclusion

Our RAG-based-LLM framework provides a generalizable, scalable solution to MRI sequence classification challenges. This method excels in resolving ambiguous cases, demonstrating potential to harmonize metadata interpretation and advance imaging informatics. Future work includes refining semantic retrieval for complex imaging tasks.

Figure(s)

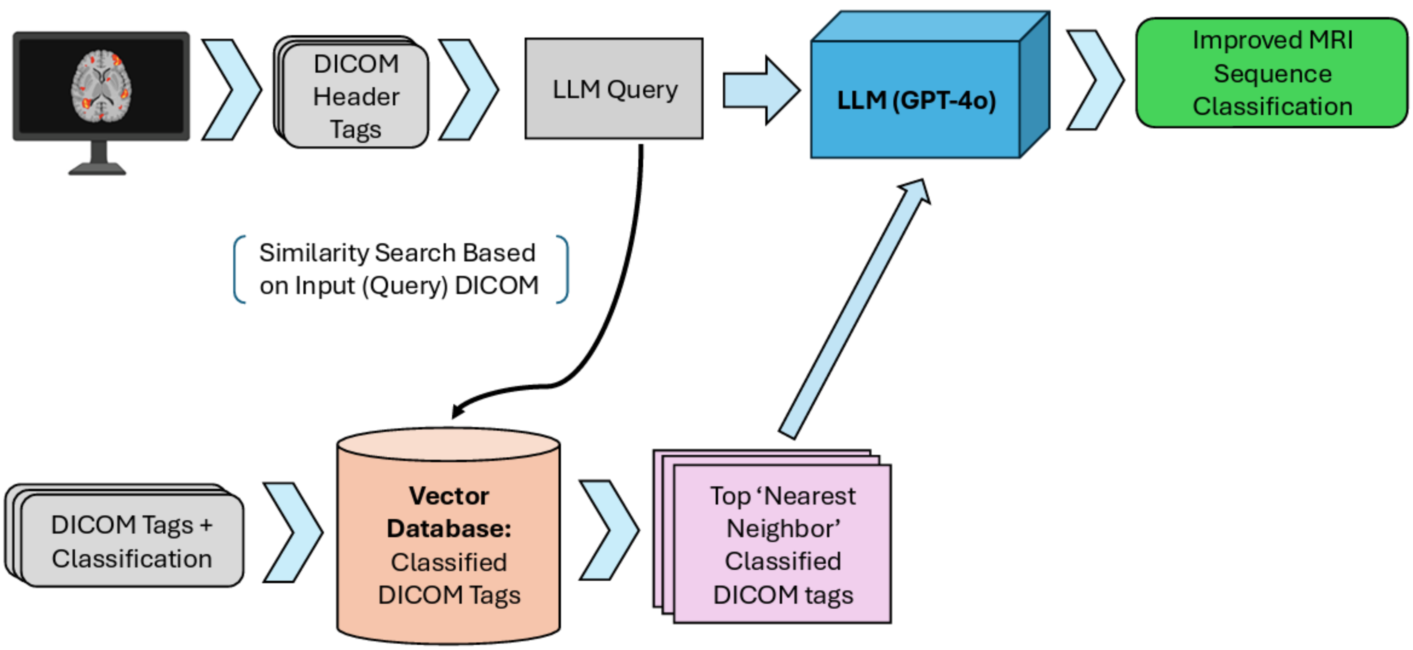

Figure 1. High-level RAG-based LLM Workflow that leverages FAISS Vector Database.

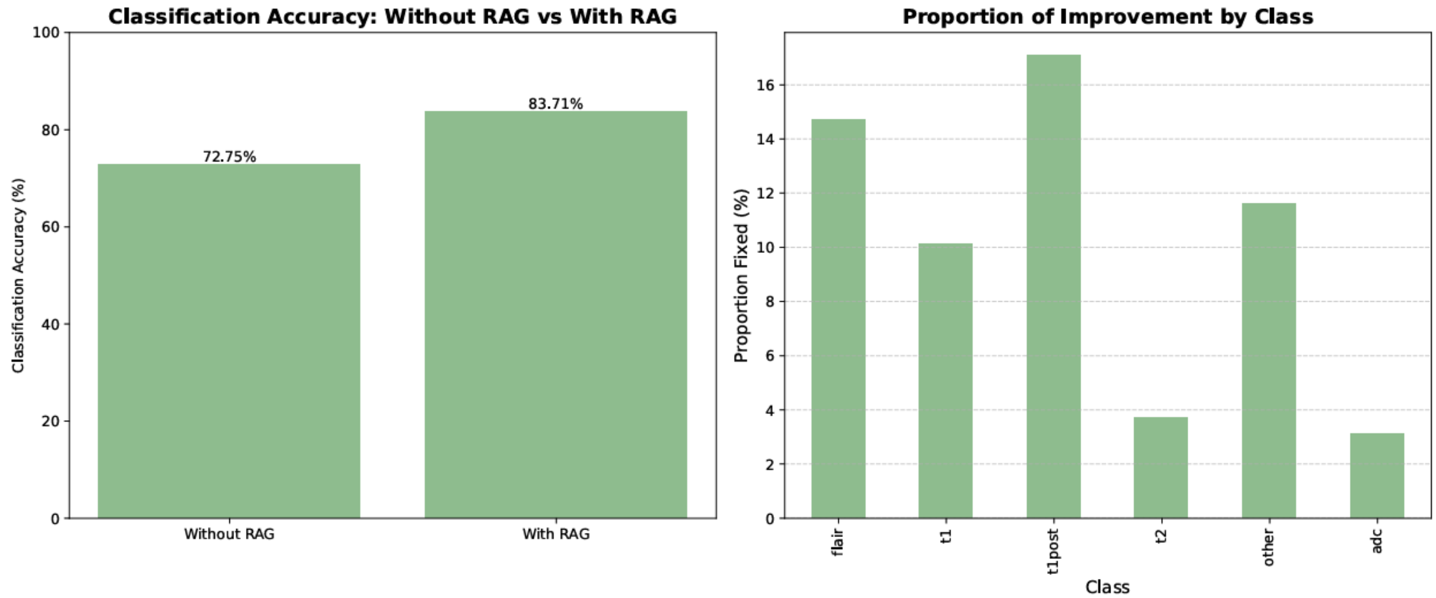

Figure 2A. (Left). MRI Sequence Classification Accuracy on Pediatric Clinical Trials brain MRIs, comparing with and without RAG-based LLM framework. RAG-based LLM framework resulted in a considerable improvement in performance (>11%). **Figure 2B** (Right). Error Reduction Rates by the RAG-Based Framework. Note that all sequences had errors reduced using RAG.

Keywords

Applications; Artificial Intelligence/Machine Learning; Emerging Technologies; Imaging Research