RAD-SRAC: Simple Retrieval Augmented Classification for Radiology

Barbara Klaudel[1, 2, 3*], **Aleksander Obuchowski**[1, 2*]

[1]TheLion.AI
[2]Polish-Japanese Academy of Information Technology
[3]Gdańsk University of Technology, Department of Decision Systems and Robotics
klaudel.b@gmail.com, obuchowskialeksander@gmail.com

Abstract

As artificial intelligence continues to advance rapidly in the healthcare sector, automated medical image analysis is increasingly used to enhance diagnostic accuracy. Large Vision Language Models (VLMs) show promise in understanding medical imagery, but their reliance on static training data often leads to outdated or inaccurate information. Current approaches to medical image classification lack the specialized understanding required for complex medical diagnostics, relying on either text-based retrieval or general-purpose image encoders. We address these limitations by developing a novel training-free retrieval-augmented classification approach that combines a specialized medical image encoder with few-shot learning across multiple imaging modalities (X-ray, CT, and MRI). Our experiments across three diverse medical imaging datasets demonstrate substantial improvements in classification performance, with F1 score gains up to 142% for state-of-the-art VLMs and 250% for smaller deployable models while requiring only 3-5 retrieved reference images, leveling the playing field for on-premise clinical applications of smaller VLMs.

Introduction

Recent advances in large language models (LLMs) have demonstrated remarkable potential across medical applications, from patient communication to clinical documentation and diagnostic support (Yang et al. 2024b), (Silverman et al. 2024), (Nakaura et al. 2024). However, their reliance on static training data often leads to outdated or inaccurate information in rapidly evolving medical fields (Omiye et al. 2024), (Sacoransky, Kwan, and Soboleski 2024). Moreover, LLMs trained on unfiltered internet data can reproduce and amplify both factual and problematic content without discrimination, leading to inconsistent outputs that range from truthful and creative to dangerously misleading (Harrer 2023). Retrieval augmented generation (RAG) (Lewis et al. 2020) has emerged as a promising solution to this limitation by dynamically incorporating external knowledge sources into the generation process. RAG architectures typically combine two key components: a retrieval mechanism that identifies relevant information from a corpus of literature or data, and a generation component that synthesizes this information into coherent, contextually appropriate responses. The application of RAG in medical imaging analysis presents unique challenges and opportunities (Gao et al. 2023). RAG systems face a delicate balance where too little context retrieval may miss crucial information while too much can introduce noise and errors, potentially degrading even originally correct model responses (Xia et al. 2024b). In clinical radiology settings, RAG systems offer particularly compelling advantages through their ability to leverage internal hospital databases (Adejumo et al. 2024) and historical patient records (Alkhalaf et al. 2024), as well as research paper archives (Jin et al. 2023) and textbooks (Wang, Ma, and Chen 2024). By retrieving similar cases, prior diagnoses, and relevant radiology reports from a hospital's own data repository, RAG-enhanced LLMs can provide context-aware assistance that reflects local patient populations and clinical practices (Wang et al. 2024). This approach not only improves diagnostic accuracy but also enables more personalized clinical decision support by incorporating institution-specific protocols, historical case outcomes or even reflecting the style of a specific radiologist (Yan et al. 2023).

While traditional LLMs excel at natural language understanding and generation, Vision Language Models (VLMs) are specifically designed to interpret and reason about visual information alongside text. There are several published research that presented VLMs tailored to the medical language and imaging (Hartsock and Rasool 2024). VLMs with RAG are particularly suitable for radiology use cases. For example, when generating radiology report impressions, RAG systems can reference similar cases from the hospital's database to ensure consistency in reporting style and terminology while maintaining high clinical efficacy metrics.

Our research extends these findings by developing a novel RAG-based approach for few-shot medical image classification. Unlike previous text-centric approaches, we implement an image-based retrieval system using the MedImageInsight (Codella et al. 2024) encoder to create specialized vector databases. Our framework retrieves visually similar medical images as few-shot examples, which are then provided to state-of-the-art VLMs including Claude 3.5 Sonnet, GPT4o, Gemini 1.5 Pro, and Qwen, as shown in Figure 1. We evaluate this approach across three diverse medical

*These authors contributed equally.

Copyright © 2025, GenAI4Health Workshop @ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

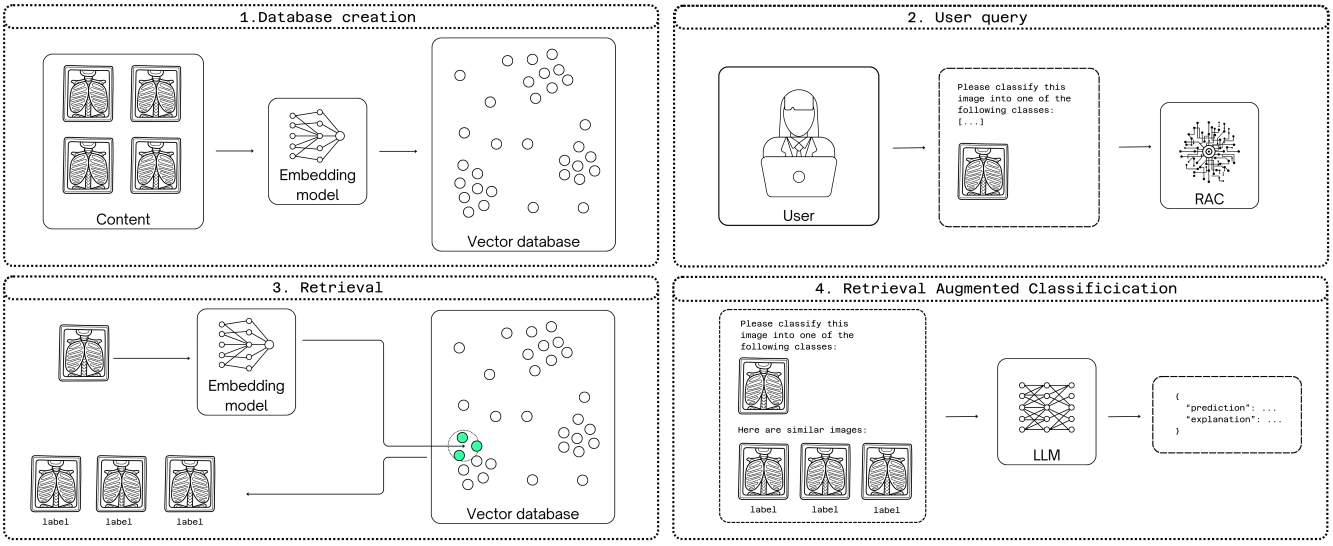

Figure 1: In the RAD-SRAC system, we first created the database of labeled images embedded with an image encoder. Then, when the user wants to use the system, similar images are retrieved from the database and appended to the user prompt. Lastly, VLM performs classification and chooses the correct label for the image.

imaging datasets: KITS23 for kidney tumors, Coronahack for chest X-rays, and Brain Tumor Classification data. This methodology requires no additional training and leverages existing datasets and models, including small LLMs, making it particularly practical for clinical applications.

The primary contributions of this study are:

1. A systematic evaluation of visual RAG's impact on medical image classification from 3 imaging modalities' performance using multiple state-of-the-art VLMs.
2. Usage of a dedicated medical image encoder, not a general purpose, unlike previous work.
3. Introduction of a simple method that does not require training and leverages available models and datasets.

Our methodology employs vector databases constructed using the medical imaging-specific encoder – MedImageInsight and general domain text decoder, facilitating efficient retrieval of relevant image features and associated metadata for augmenting LLM performance.

Code — https://github.com/TheLion-ai/RAD-SRAC

The entire code and a list of data files used for the experiments are available at the attached link. All image encoding models and datasets used to create this work are available as open source. The details of our implementation can be found in the GitHub repository.

Related Work

Recent studies have demonstrated both the potential and current limitations of LLMs in medical imaging applications (Sacoransky, Kwan, and Soboleski 2024). While LLMs show promise in understanding medical concepts and terminology, their performance varies significantly across tasks – performing better on text-based medical knowledge (80.0%) than image interpretation (45.4%) (Payne et al. 2024). Research across multiple specialties indicates that LLMs' accuracy can be enhanced through structured prompting strategies and domain-specific knowledge integration (Adams et al. 2023; Mago and Sharma 2023; Lee, Lee, and Kwon 2023; Rau et al. 2023).

Retrieval-augmented generation (RAG) has emerged as a crucial technique for reducing hallucinations (Shuster et al. 2021; Lewis et al. 2020) and enhancing LLM performance in medical imaging through various approaches. Recent work developed FactMM-RAG (Sun et al. 2024), which enhances radiology report generation through fact-aware retrieval augmentation. Their system extracts structured knowledge from radiology reports using RadGraph (Jain et al. 2021) and encodes images with MARVEL (Zhou et al. 2023) to compare with historical image-text pairs. While achieving significant improvements in both F1CheXbert (6.5%) and F1RadGraph (2%) metrics, FactMM-RAG's effectiveness is constrained by its reliance on detailed radiology reports and structured knowledge graphs.

RadioRAG (Arasteh et al. 2024) introduced a dynamic retrieval-augmented generation system for radiology diagnostics. Their approach uses GPT-3.5-turbo to extract key medical phrases from radiological questions for retrieving relevant articles from Radiopaedia.org. The system consistently enhanced diagnostic accuracy across all LLMs (improvements ranging from 2% to 54%), demonstrating RAG's potential in grounding responses in domain-specific knowledge. However, it was limited to text-only queries without direct image processing capabilities.

MMed-RAG (Xia et al. 2024a) addresses cross-domain challenges through a sophisticated architecture incorporating domain-aware retrieval mechanisms and adaptive context selection. Their experiments on the Harvard-

FairVLMed dataset revealed that naive RAG implementation can sometimes impair performance, with 43.31% of previously correct answers becoming incorrect due to retrieval interference. Their adaptive approach significantly improved the model's ability to utilize retrieved information while maintaining ground truth alignment.

Recent innovations in report generation include a novel three-component architecture (Liu et al. 2024) based on MiniGPT-4, combining visual encoding, in-domain instance induction, and coarse-to-fine decoding. Their system achieved state-of-the-art performance on IU X-RAY (Demner-Fushman et al. 2016) and MIMIC-CXR (Johnson et al. 2019) datasets through hierarchical refinement, though without using RAG during inference.

Vector databases have transformed medical image storage and retrieval, serving as specialized systems organizing data based on mathematical representations in multi-dimensional space (Pan, Wang, and Li 2024). Initial RAG implementations in medical imaging either utilized text-only vector databases (Sun et al. 2024; Arasteh et al. 2024), relied on general-purpose image encoders (Xia et al. 2024a), or trained task-specific VLMs (Liu et al. 2024).

The challenge of cross-domain knowledge transfer in medical imaging has been extensively studied (Huang et al. 2024; Tayebi Arasteh et al. 2024), particularly regarding knowledge transfer between general and specific medical domains (Mei et al. 2022). One significant advance is MedImageInsight, which demonstrated substantial improvements over general-purpose encoders in capturing domain-specific patterns and maintaining clinically relevant features.

Our research extends these findings in several key ways. We introduce a novel combination of image-based retrieval and few-shot learning for medical image classification. Unlike previous approaches that focus on single modalities, we demonstrate effectiveness across X-ray, CT, and MRI. We provide empirical evidence for the impact of domain-specific medical image retrieval on classification performance.

Table 1: The datasets used in this study.

Dataset	Modality	Classes	Count
KITS23	CT	angiomyolipoma, chromophobe_rcc, clear_cell_rcc, oncocytoma, papillary_rcc	424
Coronahack	X-ray	Normal, Pneumonia Bacteria, Pneumonia Virus	5908
Brain Tumor Classification	MRI	no_tumor, glioma_tumor, meningioma_tumor, pituitary_tumor	3264

Methods

This study used 3 open-source medical imaging datasets accessed through UMIE dataset pipelines (thelion.ai 2024): KITS23 (Heller 2024), Coronahack (Gobi 2020) (which aggregates (Cohen, Morrison, and Dao 2020) and (Kermany et al. 2018)), and Brain Tumor Classification (Bhuvaji et al. 2020) as shown in Table 1. The datasets organized patient studies differently. We decided to select a single image from each study. For KITS23, which contained multiple images per study, we selected a single representative image from each study based on the maximum tumor area visible in available segmentation masks. Coronahack and Brain Tumor classification datasets lacked study identifiers, necessitating the treatment of each image as an independent study. We extracted from each dataset 100 examples to form test sets while maintaining the original class distribution through stratification. The remaining samples were designated as database splits.

We constructed vector databases using Qdrant (Qdrant 2020) for each dataset's database split. The number of segments was set to 8 and the number of shards to 1. Image embeddings were generated utilizing the MedImageInsight encoder, creating a structured repository of image features that enabled subsequent retrieval operations. This vector database architecture formed the foundation for our simple retrieval-augmented classification (SRAC) experiments.

Our experimental design employed a structured three-part prompting protocol to systematically evaluate VLM classification performance:

System Context The system message established the model's baseline role:

```
You are a medical expert.
Analyze the {modality} image and classify
if there is a tumor present. Select the
appropriate class from {labels}.
```

Few-Shot Examples Retrieved cases were presented using one of two formats:

1. **Labeled Format:**

```
Examples:
    class: [specific_label]
    [image]
```

2. **Unlabeled Format:**

```
Examples of  similar images:
    [image]
```

In the retrieval-augmented condition, these examples were dynamically selected from our vector database using embedding similarity metrics.

Classification Request The target case was presented with a standardized instruction:

```
Now, please analyze the new image and
provide your classification.
You always need to classify.
Return only the result in JSON format:
{
```

Table 2: The precision, recall, and F1 scores of state-of-the-art VLMs with SRAC prompt enhancement with similar images with labels and without (raw).

		Claude-3-5-sonnet			**GPT-4o**			**Gemini 1.5 Pro**			**Qwen2-VL 72B Instruct**		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
KITS	Raw	0.55	0.52	0.53	0.54	0.6	0.57	0.60	0.42	0.48	0.54	0.39	0.45
	SRAC	0.57	0.67	0.61	0.62	0.64	0.63	0.57	0.60	0.59	0.58	0.6	0.59
	Gain	↑3%	↑29%	↑15%	↑15%	↑7%	↑10%	↓-4%	↑43%	↑24%	↑10%	↑54%	↑31%
Coronahack	Raw	0.47	0.47	0.46	0.41	0.42	0.41	0.39	0.33	0.28	0.49	0.3	0.35
	SRAC	0.78	0.76	0.76	0.76	0.76	0.76	0.72	0.68	0.67	0.6	0.5	0.5
	Gain	↑66%	↑62%	↑66%	↑87%	↑81%	↑85%	↑84%	↑106%	↑142%	↑22%	↑67%	↑44%
Brain Tumor	Raw	0.60	0.56	0.56	0.65	0.58	0.59	0.73	0.51	0.44	0.486	0.3	0.35
	SRAC	0.92	0.91	0.91	0.94	0.94	0.94	0.92	0.91	0.91	0.698	0.6	0.61
	Gain	↑52%	↑63%	↑62%	↑46%	↑62%	↑59%	↑26%	↑78%	↑107%	↑44%	↑100%	↑76%

```
    'y_pred': predicted_class,
    'explanation': brief_explanation
}
```

This prompting framework was consistently applied across all evaluated VLMs to ensure a fair comparison, with the only variation being the presence or absence of a second prompt for SRAC versus baseline conditions. The structured output format enabled systematic evaluation of model predictions while capturing the reasoning process through brief explanations, providing both quantitative classification results and qualitative insight into the model's diagnostic approach.

Experiments and Results

In the first experiment, we evaluated the impact of SRAC on the image classification performance of the state-of-the-art VLMs (according to the OpenVLM leaderboard (Duan et al. 2024)) through a controlled comparison. The control condition involved the direct presentation of test split images to Claude 3.5 Sonnet (Anthropic 2024), GPT4o (Hurst et al. 2024), Gemini 1.5 Pro (Team et al. 2024) and Qwen2-VL 72B (Yang et al. 2024a) VLMs for classification in Table 2. For each of the models, temperature was set to 1.

In the SRAC implementation, we enhanced this process by first retrieving the five images with the smallest dot product distance from the corresponding dataset's vector database. These retrieved images and associated labels were provided as few-shot examples to the LLMs during the classification task. For each of the experiments, if the VLM's response contained errors, e.g., not following structured input, we rerun the experiment for this image up to 5 times.

For the "raw" experiments, without few-shot examples, we used only prompts 1 and 3, while for the "SRAC" experiments we used all of the prompt messages and supplied the model with 5 images with labels, before asking to classify a new image. The experiments demonstrate substantial improvements across all three leading VLMs when enhancing prompts with similar images. The most significant gains were observed in the Coronahack dataset, where SRAC enhancement led to performance improvements ranging from 66% to 142% across all metrics. Claude-3-5-sonnet showed the most consistent performance, achieving balanced precision and recall improvements, particularly in the Brain Tumor dataset where both metrics exceeded 0.90 after SRAC enhancement. While GPT-4 maintained competitive performance, achieving the highest F1 score of 0.94 on the Brain Tumor dataset, Gemini 1.5 Pro exhibited the most variable results, showing the largest gains in recall but sometimes at the cost of precision, as evidenced by the 4% precision decrease in the KITS dataset. Notably, all models demonstrated more modest improvements on the KITS23 dataset compared to other tasks.

We also tested the impact of showing labeled examples to the model vs. presenting the most similar examples without the labels to the model. For this purpose, we used the same models as in the previous experiment in Table 3. The experimental results reveal varying performance across the three VLMs on different medical imaging datasets. Unlabeled few-shot prompting had varied effects on the performance. Claude's performance decreased for CoronaHack and Brain Tumor Classification compared with the "raw" experiment, while for KITS-23 it remained relatively the same. For GPT 4o, the performance on KITS-23 decreased, while for the other 2 datasets, it has significantly improved. Gemini's performance improved on all 3 datasets.

Due to stringent healthcare data protection regulations, patient data often cannot be transmitted beyond hospital internal networks, which precludes the use of commercial LLMs such as Claude or OpenAI's offerings that require data processing on external company servers. To address this common constraint in healthcare settings, we evaluated our solution using smaller models that could be deployed on-premise within hospital infrastructure, thereby maintaining data privacy by processing sensitive patient information exclusively within the institution's secure environment. For this purpose, we evaluated the "raw" vs "RAG" performance of Gemini 1.5 Flash-8B and Pixtral-12B (Agrawal et al. 2024) with only 8 and 12 billion parameters respectively in Table 4. Although Gemini 1.5 Flash is not open source, it is a small model and, therefore, a good indicator of the performance of smaller LLMs.

The results demonstrate that SRAC significantly improves the performance of both smaller models, with particularly striking improvements for the Pixtral-12B model. Most notably, Pixtral-12B showed remarkable gains on the Coronahack dataset, with a 250% improvement in F1 score, and similarly impressive gains on the Brain Tumor dataset with increases of 194% in precision and 162% in recall. The

Table 3: The precision, recall, and F1 scores of state-of-the-art VLMs with SRAC prompt enhancement – images without labels.

		Claude-3-5-sonnet			**GPT 4o**			**Gemini 1.5 Pro**			**Qwen2-VL 72B Instruct**		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
KITS	SRAC unlabelled	0.57	0.50	0.53	0.57	0.52	0.54	0.56	0.47	0.51	0.54	0.43	0.48
Coronahack	SRAC unlabelled	0.37	0.37	0.37	0.43	0.46	0.42	0.46	0.38	0.38	0.34	0.3	0.3
Brain Tumor	SRAC unlabelled	0.60	0.46	0.45	0.65	0.60	0.60	0.73	0.63	0.60	0.41	0.4	0.39

Table 4: The precision, recall, and F1 scores of small state-of-the-art VLMs with SRAC prompt enhancement with similar images with labels and without (raw).

		Gemini 1.5 Flash-8B			**Pixtral-12B**		
		P	R	F1	P	R	F1
KITS	Raw	0.54	0.4	0.45	0.51	0.38	0.43
	SRAC	0.56	0.72	0.63	0.51	0.54	0.52
	Gain	↑3%	↑80%	↑40%	↓1%	↑42%	↑21%
Corona-hack	Raw	0.34	0.46	0.39	0.22	0.26	0.17
	SRAC	0.47	0.51	0.43	0.57	0.60	0.58
	Gain	↑38%	↑10%	↑9%	↑163%	↑131%	↑240%
Brain Tumor	Raw	0.72	0.51	0.44	0.25	0.26	0.23
	SRAC	0.72	0.68	0.66	0.72	0.68	0.66
	Gain	0%	↑33%	↑51%	↑194%	↑162%	184%

Gemini 1.5 Flash-8B model exhibited more modest but still significant improvements, with its most substantial gain being an 80% increase in recall on the KITS dataset. Interestingly, both models achieved identical final performance on the Brain Tumor dataset after SRAC enhancement, reaching a precision of 0.72 and recall of 0.68, despite Pixtral-12b starting from a much lower baseline. This suggests that SRAC can potentially level the playing field between models of different sizes, particularly for specialized medical imaging tasks.

We also tested the optimal number of examples to show to VLM before classification. The results shown in Table 2 reveal that increasing the number of reference images does not linearly correlate with improved performance across the tested models on the KITS23 dataset. Claude-3-5-Sonnet demonstrates the most stable performance pattern, reaching its peak F1 score of 0.61 with just 3 reference images and maintaining consistent performance even up to 10 images. GPT-4o shows significant improvement from 1 to 3 images, achieving its optimal performance of 0.628 with 3 images, but experiences a decline in performance with additional images beyond that point. Gemini 1.5 Pro exhibits more variable behavior, with performance fluctuating across different image counts and reaching its peak F1 score of 0.59 at 5 images. Notably, all models show substantial improvement when moving from 1 to 3 reference images, suggesting that even a minimal SRAC enhancement provides meaningful benefits. This finding has practical implications for real-world applications, as it suggests that using 3-5 reference images may be optimal for balancing performance gains with computational efficiency.

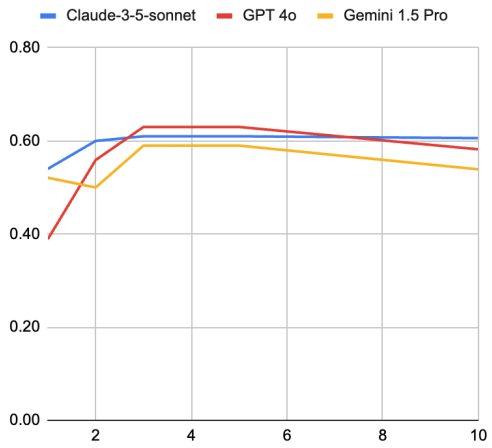

Figure 2: F1 scores of VLMs across different numbers of reference images (1-10), showing optimal performance at 3-5 images.

Discussion

The substantial variation in SRAC's effectiveness across different datasets and models provides important insights into the method's strengths and limitations. The dramatic improvements observed in Coronahack (66-142% gain) and Brain Tumor Classification (62-107% gain) datasets contrast sharply with the more modest gains in KITS23 (10-22% gain). This disparity stems from the relative complexity of the classification problems. Brain Tumor Classification is the simplest problem since the 3 possible tumor classes occur in different brain regions (Osborn, Hedlund, and Salzman 2017). Coronahack presents the more difficult problem of pneumonia classification, which is also toughened by the fact that the X-rays present children. Radiologists do not diagnose pneumonia causative organisms since it is believed that generally, CXR shadows do not allow distinguishing between viral and bacterial pneumonia (De Lacey, Morley, and Berman 2012). However, it is a popular applied deep learning problem with many successful approaches (Sharma and Guleria 2024) achieving accuracies up to 99.61% on this dataset (Gour and Jain 2022). KITS-23's kidney tumor classification is a significantly more challenging problem, that poses a real diagnostic challenge, with small benign tumors

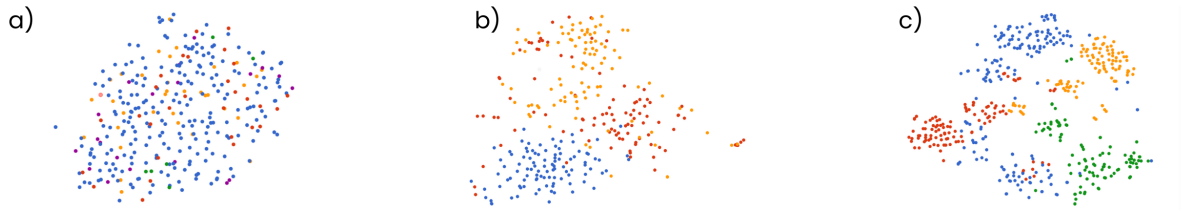

Figure 3: T-sne visualization of the image embeddings for each dataset: a) KITS, b) Coronahack, c) Brain Tumor

often misdiagnosed as malignant and found to be benign only after postoperative histopathology examination (Kay and Pedrosa 2016).

The difficulty of the problems was also visible in the t-SNE visualizations of the datasets (shown in Figure 3.) The visualizations for Coronahack and Brain Tumor Classification show a clear separation between the classes, forming clusters. For KITS-23 the classes are scattered and the overbalance of clear cell renal carcinoma samples is evident. This is in line with the clinical practice, where it is the most commonly occurring tumor. The effectiveness of our SRAC approach is fundamentally tied to the quality of the MedImageInsight encoder's representations. The more modest improvements in KITS23 classification suggest that the encoder may not fully capture the subtle distinctions between different types of kidney tumors. This limitation points to a crucial area for future development: the need for more sophisticated medical image encoders that can better capture fine-grained pathological features.

Our investigation into the optimal number of reference images revealed a consistent pattern across all tested models: the relationship between the number of reference images and classification performance follows a clear diminishing returns curve. The finding that 3-5 reference images typically yield optimal performance has significant practical implications. This "sweet spot" suggests that models can effectively learn from a small set of well-chosen examples without requiring extensive computational resources or suffering from information overload.

The strong performance improvements observed in smaller models like Pixtral-12B (up to 250% F1 score improvement) have important implications for clinical deployment. These results suggest that SRAC can enable on-premise deployments that meet both performance requirements and privacy constraints. The ability to achieve competitive performance with smaller models addresses a crucial barrier to adoption in healthcare settings where data cannot be transmitted to external servers. Implementation considerations must account for computational resources required for maintaining and querying vector databases, integration with existing PACS (Picture Archiving and Communication Systems) and clinical workflows, real-time performance requirements for clinical decision support, and data privacy and security compliance.

Several limitations of our current approach suggest directions for future research. The current MedImageInsight encoder's performance on complex cases like KITS23 indicates the need for more sophisticated medical imaging-specific encoders, which capture subtle pathological features better. While our study covered three different imaging modalities, expanding to a broader range of pathologies and modalities would provide a more comprehensive validation of the approach. Investigation of cross-domain knowledge transfer through alternative vector database contents could reveal opportunities for leveraging broader medical imaging knowledge bases.

Our findings suggest that SRAC has the potential to significantly impact medical imaging workflows by improving the accuracy of automated image classification, particularly for smaller deployable models. The ability to leverage existing image databases for improved classification could help institutions better utilize their historical data for current diagnostic challenges. Furthermore, the approach could serve as a valuable educational tool, automatically providing relevant reference cases during training and continuing education. The demonstrated effectiveness of smaller models with SRAC enhancement suggests a path toward more widespread deployment of AI assistance in medical imaging, even in resource-constrained settings. The results of SRAC-enhanced small LLMs almost matched the results of raw LLMs, suggesting that the performance of small LLMs can be significantly boosted to match the performance of larger models with SRAC. Small LLMs are key for healthcare implementations since such models can be deployed locally without the need to share the data with the external provider.

Conclusions

This study demonstrates the significant potential of Retrieval-Augmented Classification in enhancing medical image classification across multiple imaging modalities. Our experiments revealed substantial improvements in classification performance when augmenting VLMs with retrieved similar cases, with gains ranging from 10% to 250% in F1 scores across different datasets and models. The finding that optimal performance can be achieved with just 3-5 reference images, combined with the strong performance of smaller, deployable models, indicates the practical viability of this approach in clinical settings where data privacy is paramount.

References

Adams, L. C.; Truhn, D.; Busch, F.; Kader, A.; Niehues, S. M.; Makowski, M. R.; and Bressem, K. K. 2023. Leveraging GPT-4 for post hoc transformation of free-text radiology

reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4): e230725.

Adejumo, P.; Thangaraj, P. M.; Vasisht Shankar, S.; Dhingra, L. S.; Aminorroaya, A.; and Khera, R. 2024. Retrieval-Augmented Generation for Extracting CHA2DS2VASc Features from Unstructured Clinical Notes in Patients with Atrial Fibrillation. *medRxiv*, 2024–09.

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Chaplot, D.; Chudnovsky, J.; Garg, S.; Gervet, T.; Ghosh, S.; Héliou, A.; Jacob, P.; et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Alkhalaf, M.; Yu, P.; Yin, M.; and Deng, C. 2024. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, 104662.

Anthropic, A. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*.

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

Bhuvaji, S.; Kadam, A.; Bhumkar, P.; Dedge, S.; and Kanchan, S. 2020. Brain Tumor Classification (MRI).

Codella, N. C.; Jin, Y.; Jain, S.; Gu, Y.; Lee, H. H.; Abacha, A. B.; Santamaria-Pang, A.; Guyman, W.; Sangani, N.; Zhang, S.; et al. 2024. MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging. *arXiv preprint arXiv:2410.06542*.

Cohen, J. P.; Morrison, P.; and Dao, L. 2020. COVID-19 image data collection. *arXiv 2003.11597*.

De Lacey, G.; Morley, S.; and Berman, L. 2012. *The chest X-ray: a survival guide*. Elsevier Health Sciences.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; Lin, D.; and Chen, K. 2024. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. arXiv:2407.11691.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gobi, P. 2020. CoronaHack -Chest X-Ray-Dataset.

Gour, M.; and Jain, S. 2022. Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification. *Computers in biology and medicine*, 140: 105047.

Harrer, S. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90.

Hartsock, I.; and Rasool, G. 2024. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*.

Heller, N. 2024. *Kidney and Kidney Tumor Segmentation: MICCAI 2023 Challenge, KiTS 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings*, volume 14540. Springer Nature.

Huang, Y.; Zou, J.; Meng, L.; Yue, X.; Zhao, Q.; Li, J.; Song, C.; Jimenez, G.; Li, S.; and Fu, G. 2024. Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification. *arXiv preprint arXiv:2402.07595*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Jin, Q.; Kim, W.; Chen, Q.; Comeau, D. C.; Yeganova, L.; Wilbur, W. J.; and Lu, Z. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11): btad651.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kay, F. U.; and Pedrosa, I. 2016. Imaging of solid renal masses. *Radiologic Clinics of North America*, 55(2): 243.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5): 1122–1131.

Lee, K. H.; Lee, R. W.; and Kwon, Y. E. 2023. Validation of a Deep Learning Chest X-ray Interpretation Model: Integrating Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT. *Diagnostics*, 14(1): 90.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18635–18643.

Mago, J.; and Sharma, M. 2023. The potential usefulness of ChatGPT in oral and maxillofacial radiology. *Cureus*, 15(7).

Mei, X.; Liu, Z.; Robson, P. M.; Marinelli, B.; Huang, M.; Doshi, A.; Jacobi, A.; Cao, C.; Link, K. E.; Yang, T.; et al.

2022. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315.

Nakaura, T.; Yoshida, N.; Kobayashi, N.; Shiraishi, K.; Nagayama, Y.; Uetani, H.; Kidoh, M.; Hokamura, M.; Funama, Y.; and Hirai, T. 2024. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2): 190–200.

Omiye, J. A.; Gui, H.; Rezaei, S. J.; Zou, J.; and Daneshjou, R. 2024. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of Internal Medicine*, 177(2): 210–220.

Osborn, A. G.; Hedlund, G. L.; and Salzman, K. L. 2017. Osborn's brain e-book.

Pan, J. J.; Wang, J.; and Li, G. 2024. Survey of vector database management systems. *The VLDB Journal*, 33(5): 1591–1615.

Payne, D. L.; Purohit, K.; Borrero, W. M.; Chung, K.; Hao, M.; Mpoy, M.; Jin, M.; Prasanna, P.; and Hill, V. 2024. Performance of GPT-4 on the American College of Radiology In-training Examination: Evaluating Accuracy, Model Drift, and Fine-tuning. *Academic radiology*.

Qdrant. 2020. Qdrant.

Rau, A.; Rau, S.; Zoeller, D.; Fink, A.; Tran, H.; Wilpert, C.; Nattenmueller, J.; Neubauer, J.; Bamberg, F.; Reisert, M.; et al. 2023. A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology*, 308(1): e230970.

Sacoransky, E.; Kwan, B. Y.; and Soboleski, D. 2024. ChatGPT and assistive AI in structured radiology reporting: A systematic review. *Current Problems in Diagnostic Radiology*.

Sharma, S.; and Guleria, K. 2024. A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images. *Multimedia Tools and Applications*, 83(8): 24101–24151.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Silverman, A. L.; Sushil, M.; Bhasuran, B.; Ludwig, D.; Buchanan, J.; Racz, R.; Parakala, M.; El-Kamary, S.; Ahima, O.; Belov, A.; et al. 2024. Algorithmic Identification of Treatment-Emergent Adverse Events From Clinical Notes Using Large Language Models: A Pilot Study in Inflammatory Bowel Disease. *Clinical Pharmacology & Therapeutics*, 115(6): 1391–1399.

Sun, L.; Zhao, J.; Han, M.; and Xiong, C. 2024. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*.

Tayebi Arasteh, S.; Misera, L.; Kather, J. N.; Truhn, D.; and Nebelung, S. 2024. Enhancing diagnostic deep learning via self-supervised pretraining on large-scale, unlabeled non-medical images. *European Radiology Experimental*, 8(1): 10.

Team, G.; Reid, M.; Savinov, N.; Teplyashin, D.; Dmitry, L.; Lillicrap, T.; Alayrac, J.; Soricut, R.; Lazaridou, A.; Firat, O.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. In arXiv [cs. CL]. arXiv.

thelion.ai. 2024. umie_datasets (Revision 5384587).

Wang, G.; Ran, J.; Tang, R.; Chang, C.-Y.; Chuang, Y.-N.; Liu, Z.; Braverman, V.; Liu, Z.; and Hu, X. 2024. Assessing and enhancing large language models in rare disease question-answering. *arXiv preprint arXiv:2408.08422*.

Wang, Y.; Ma, X.; and Chen, W. 2024. Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1754–1770. Miami, Florida, USA: Association for Computational Linguistics.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Xia, P.; Zhu, K.; Li, H.; Zhu, H.; Li, Y.; Li, G.; Zhang, L.; and Yao, H. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1081–1093.

Yan, B.; Liu, R.; Kuo, D.; Adithan, S.; Reis, E.; Kwak, S.; Venugopal, V.; O'Connell, C.; Saenz, A.; Rajpurkar, P.; and Moor, M. 2023. Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14676–14688. Singapore: Association for Computational Linguistics.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yang, Z.; Xu, X.; Yao, B.; Rogers, E.; Zhang, S.; Intille, S.; Shara, N.; Gao, G. G.; and Wang, D. 2024b. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2): 1–35.

Zhou, T.; Mei, S.; Li, X.; Liu, Z.; Xiong, C.; Liu, Z.; Gu, Y.; and Yu, G. 2023. Unlock Multi-Modal Capability of Dense Retrieval via Visual Module Plugin. *arXiv preprint arXiv:2310.14037*.