Towards Interactive and Interpretable Image Retrieval-Based Diagnosis: Enhancing Brain Tumor Classification with LLM Explanations and Latent Structure Preservation

Pranav Manjunath(✉), Brian Lerner, and Timothy Dunn

Duke University, Durham, NC 27005, USA
{pranav.manjunath,timothy.dunn}@duke.edu

Abstract. When evaluating patient scans, clinicians make use of their previous experience to make a diagnosis. However, for complex conditions such as brain tumors, the availability of relevant information beyond a clinician's personal context becomes more valuable. While the application of content-based image retrieval (CBIR) for medical images is not new, their introduction into practice has been relatively minimal. Where older CBIR systems relying on manually extracted image features suffered from poor performance, newer systems incorporating deep learning may have better performance but decreased interpretability. In this study, we present an interactive image retrieval system that enables accurate and interpretable brain tumor classification. We show that image encoders trained with supervised contrastive learning preserve latent structure within the retrieval space and exhibit classification performance on par with, or exceeding, that of conventional black box classifiers. We integrate off-the-shelf LLMs to enhance the system's accessibility through retrieval report summarization and user Q&A interactions. We recognize the importance of developing clinician-ML systems by providing a framework that clinicians can not only trust the performance of but can interact with. Our findings provide for the seed of a system that can augment the performance of human clinicians in a process that mirrors their natural thought patterns, while increasing the speed of their interactions with medical CBIR through LLMs.

Keywords: Medical CBIR · Brain Tumor Classification · Large Language Models · Deep Learning · Contrastive Learning

1 Introduction

While brain tumors rarely occur, when they do, they are one of the most deadly forms of cancer [1]. Magnetic resonance imaging (MRI) is an important modality employed to identify tumors, but disease complexity and overlapping phenotypes can slow or stymie diagnosis [2]. Machine learning (ML) has proven capable of solving a wide variety of important healthcare problems, but its clinical adoption has been slow [3], despite

J. Finkelstein et al. (Eds.): AIME 2024, LNAI 14845, pp. 335–349, 2024.
https://doi.org/10.1007/978-3-031-66535-6_35

evidence that it can often surpass the performance of human clinicians in the task of classifying diseases in medical imaging [4–6]. Hampered by a distrust of ML, attempts to translate research into the clinic [7] have been sparse. Conventional black box ML models usually fail to offer human-interpretable insight into their decision-making processes–a flaw poorly tolerated by patients, clinicians, and administrators alike. The implementation of interpretable frameworks remain essential to clinicians in cancer diagnosis [8]. When clinicians engage in diagnosis, they often relate past encounters to the present one–behavior that is reminiscent of content-based image retrieval (CBIR). Previous work has shown that retrieved images which are structurally like the query image instill more trust in the user [9]. In certain contexts, images are naturally accompanied by detailed descriptions, i.e. radiology reports.

In this study, we design and implement an interactive LLM-explained CBIR system that is interpretable and offers high performance in brain tumor classification (Fig. 1). The system follows a two-step process: i) it employs a deep-learning based image encoder to identify scans similar to the query scan from a vector database (CBIR) favoring latent structure preservation ii) it classifies the query image by running a neighborhood classifier on these retrieved images, thereby enhancing example-based interpretability. We improve upon neighborhood classifiers by introducing a hyperparameter-free version, nearest incorrect retrieval (NIR). To enable interactivity, we pioneer an approach that utilizes off-the-shelf LLMs to provide summarization and dynamic Q&A over text descriptions of retrieved scans.

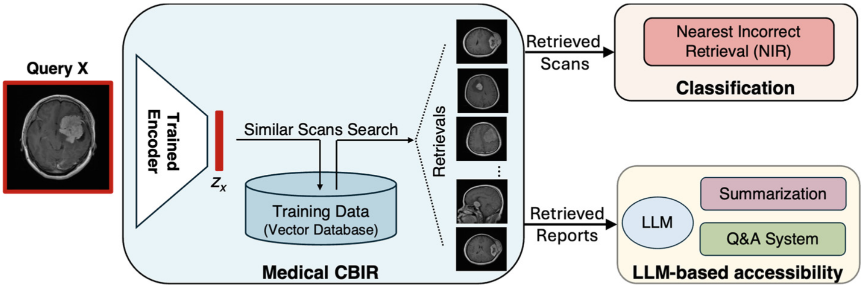


Fig. 1. The process of LLM-explained CBIR for interpretable medical image classification. A query image is encoded into z_x by a deep learning-based encoder. z_x is used to search the training database for similar images and retrieve them. The retrievals are used for downstream tasks like classification, while text descriptions (the report) of each retrieved image are passed to an LLM for summarization and question answering.

2 Related Work

CBIR has origins predating deep-learning algorithms, yet the advent of deep learning has significantly enhanced the effectiveness of CBIR in the medical domain [9, 10]. As we do here, previous work has framed CBIR in terms of example-based interpretability and posited that CBIR-based classification can instill trust in users [11]. [12, 13] trained a deep-learning based encoder to classify and retrieve various modalities of medical imaging, though in contrast to our approach they focus on training with cross-entropy (CE). Other work has explored using contrastive learning to train models for medical image retrieval [11, 14]. Here, we evaluate retrieval and classification using both CE and contrastive learning.

To provide explanations on why certain images are retrieved for a given query, [11] explore saliency-based explanations between the query and the retrieved images for chest X-rays. However, saliency maps have been shown to be inconsistent and untrustworthy for medical imaging [15]. Our approach focuses on creating inherent interpretability through CBIR, which is in turn made more accessible with LLMs.

3 Data

We utilize a brain tumor dataset [16] consisting of 3064 T1-weighted and contrast-enhanced 2D MRI slices. These slices span 233 patients, with a mixture of axial, coronal, and sagittal slices attributed to each. Each slice contains a segmented tumor that falls into one of three categories. Followed by the respective number of slices, these categories are meningioma (708), glioma (1426), and pituitary (930). Plane distribution within each tumor type is roughly uniform. We use train and test partitions from [16], comprising 186 and 47 unique patients respectively.

4 Preservation of Latent Structure in CBIR

We define latent structure as those aspects of an image which are relevant to visual understanding but are not explicitly used while training. As structural similarities in retrieved images correlates with user trust [9], we use the known values of anatomical plane and tumor size as a proxy for latent structural properties and hold that latent structure is preserved when these properties are highly similar for a query image and its top retrievals. For anatomical plane, we measure the fraction of retrievals that match the plane of the query and denote this as anatomical consistency. For tumor size, we measure the average size difference between query and retrievals. Latent structure is investigated for various training objectives.

5 Neighborhood Classification

As an alternative to K-Nearest Neighbors (KNN), we propose Nearest Incorrect Retrieval (NIR), which does not need hyperparameter tuning and is less likely to overfit on the validation set (Appendix 1). To get the NIR, we sort a query's retrievals in order of decreasing distance. Using c_n, the array of retrieval class labels, for each class c we obtain I, the index of the first neighbor that is not of class c. We store I for each class in I_c an array of $[1xC]$ where C is the total number of classes.

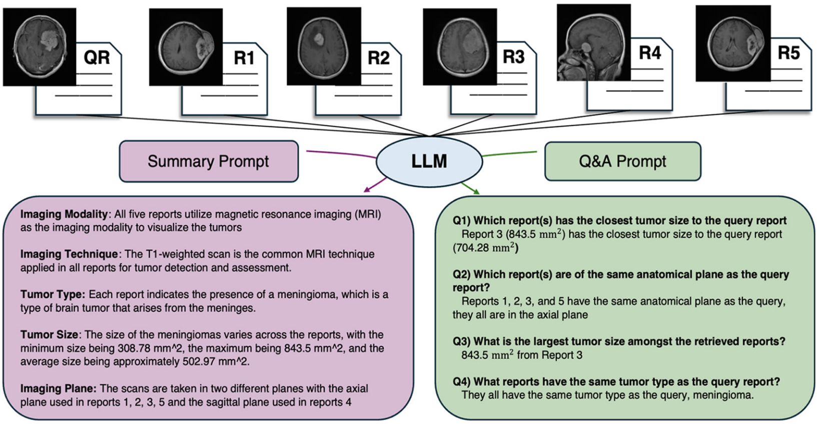
$$I_c = \left[argmin(c_n \neq c) | c \in C\right] \tag{1}$$

We convert I_c into probabilities using a softmax function σ and assign the class with the highest probability as the query class Q.

$$Q_c = argmax(\sigma(I_c)) \tag{2}$$

In contrast to KNN, the number of neighbors that NIR incorporates into its process is not bounded by a set number. Knowing I indicates how far within the neighborhood space the algorithm has to look to obtain the first incorrect neighbor. Thus, I can be considered a measure of model confidence. We also introduce Iterative-NIR (I-NIR), a variant of NIR that has a higher tolerance to error, offering a more lenient evaluation framework compared to the original NIR metric. We present results for both the standard NIR and I-NIR, with the latter examining the 10th nearest incorrect retrieval.

6 LLM-Explained CBIR



Imaging Modality: All five reports utilize magnetic resonance imaging (MRI) as the imaging modality to visualize the tumors

Imaging Technique: The T1-weighted scan is the common MRI technique applied in all reports for tumor detection and assessment.

Tumor Type: Each report indicates the presence of a meningioma, which is a type of brain tumor that arises from the meninges.

Tumor Size: The size of the meningiomas varies across the reports, with the minimum size being 308.78 mm^2, the maximum being 843.5 mm^2, and the average size being approximately 502.97 mm^2.

Imaging Plane: The scans are taken in two different planes with the axial plane used in reports 1, 2, 3, 5 and the sagittal plane used in reports 4

Q1) Which report(s) has the closest tumor size to the query report
Report 3 (843.5 mm^2) has the closest tumor size to the query report (704.28 mm^2)

Q2) Which report(s) are of the same anatomical plane as the query report?
Reports 1, 2, 3, and 5 have the same anatomical plane as the query, they all are in the axial plane

Q3) What is the largest tumor size amongst the retrieved reports?
843.5 mm^2 from Report 3

Q4) What reports have the same tumor type as the query report?
They all have the same tumor type as the query, meningioma.

Fig. 2. Integration of LLM in CBIR for medical text summary and Q&A system. QR indicates the query report and R1–5 are the reports for each of the retrieved images. Summarization uses retrieved reports while Q&A uses both the query and retrieved report.

6.1 Image Description Generation

As there are no radiology reports present in the original dataset, we generate a pseudo-text description for each scan by extracting the following attributes: imaging modality, imaging technique, tumor size, tumor type, and anatomical plane. In this paper, we refer to these text descriptions as “reports”. We generated three distinct report types of varying complexity to test the robustness of LLM interpretation which we use for the.

LLM-based summarization and Q&A tasks. Descriptions of examples are in Appendix 2.

1. Plain Report: Hardcoded sentence structure; reports only differ based on the extracted values of the respective scan.
2. Normal Report: Generated by GPT-4, this report gives a concise breakdown of the attributes roughly in the style of a radiologist. Natural variation in style occurs due to the nature of LLMs.
3. Alternative Report: This report takes the normal report and randomly performs the following swaps for roughly half of the reports: tumor size units ($\text{mm}^2 \rightarrow \text{cm}^2$), tumor type (meningioma → meningeal tumor, glioma → glial tumor, pituitary tumor → pituitary gland tumor), and anatomical plane (axial → transverse, sagittal → longitudinal, coronal → frontal).

6.2 Summarization

Summarization of query's retrievals is a common task in CBIR that is conventionally handled by the human user. Here we evaluate the capability of an LLM for generating a summary that is accurate, relevant and complete (RC), these metrics are explained in Appendix 4.2. After reports are generated for all scans, 50 sets containing the report of a query scan and those for the 5 nearest neighbors are fed into the LLM, which is prompted to generate a detailed and comprehensive summary (Fig. 2).

6.3 Interactive Question and Answer (Q&A) System

We also develop and probe an LLM-based Q&A system, designed to simulate interactivity with the retrieved data. We supply the LLM with both the query report and the top five retrieved reports. Subsequently, the LLM is prompted to answer questions that pertain directly to the content of these reports (Fig. 2). An interactive Q&A system allows users to investigate the specific aspects of reports that are most relevant to their needs. The questions asked in this experiment are:

1. Which report(s) has the closest tumor size to the query report?
2. Which report(s) are of the same anatomical plane as the query report?
3. What is the largest tumor size amongst the retrieved reports?
4. What reports have the same tumor type as the query report?

7 Experiment Design

7.1 Training Objectives

To determine the paradigm that maximizes both latent structure and classification performance, we train image encoders using a range of training objectives. We employ four different contrastive losses: triplet loss [17], Neighborhood Component Analysis (NCA) [18], InfoNCE [19], and Supervised Contrastive Loss (SupCon) [20]. All of these are supervised except for InfoNCE, which is essentially an unsupervised form of SupCon that we use to gauge the impact of supervision. We compare these contrastive losses with black-box models trained end-to-end with CE.

7.2 Image Encoders

We run the described training objectives on two CNN-based encoder architectures: EfficientNet-B2 (EN-B2) as [21] shows the B2 variant has better performance on this dataset, and ResNet50. Both use pretrained ImageNet weights and output a 128-length embedding used for downstream tasks. For CE trained models, we train an MLP on top of the embedding layer, after which we perform classification using i) the class probabilities output from the MLP, i.e. black box, and ii) the output of the embedding layer. Data augmentation and further training details are discussed in Appendix 5.

7.3 Evaluation

Query images from the test set are classified using NIR, I-NIR, and KNN. For KNN, we focus on K values of 5, 7, and 25, which fall within the upper bound that can be cognitively interpreted by humans [22]. For a given query image, similar images are retrieved based on the Euclidean distance between their embeddings and the query embedding. Classification performance is measured by AUROC and Kappa statistic (average and ensemble) and the retrieval space is assessed for anatomical consistency and average tumor size.

7.4 LLM

We query OpenAI's gpt-4–1106-preview (GPT-4) and gpt-3.5-turbo-1106 (GPT-3.5), using the same set of prompts (Appendix 3) for inference. We prioritize the use of off-the-shelf LLMs due to their ready-to-use nature and ease of accessibility, eliminating the need for fine-tuning. Report generation, summarization, and Q&A tasks are validated on 50 sets of retrievals and assessed for accuracy. Summarization is also assessed for RC. The data utilized for this project is open source, contains no PHI, and is thus safe for use with these LLMs.

8 Results

Given the superior classification performance of EN-B2 vs. ResNet50, only the former is used to generate the following results. Results for both encoders across training objectives on the classification task are shown in Appendix 6.

8.1 Retrieval

Figure 3 illustrates that SupCon preserves higher latent structure when compared to other supervised training objectives in both anatomical plane and average tumor size difference. However, we notice that InfoNCE preserves the highest latent structure when compared to all training objectives for both metrics. The non-contrastive CE preserves the least latent structure.

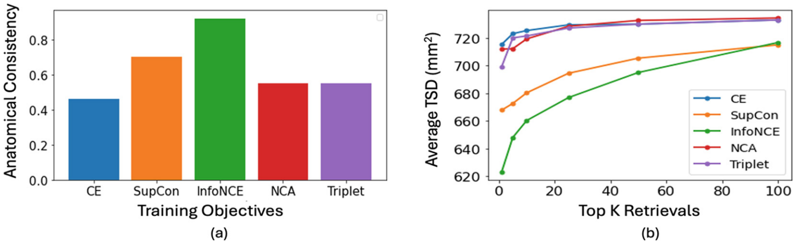


Fig. 3. Latent structure preservation results. (a): Anatomical consistency for top 5 retrievals. (b) Average tumor size difference (TSD) for top K retrievals

8.2 Classification

In Table 1, we demonstrate that NIR and I-NIR achieve comparable classification performance to KNN, and slightly exceed it for SupCon, triplet, and CE. While all training objectives are comparable for AUROC, SupCon shows a superior Kappa of 0.91. This is important because Kappa offers greater robustness to class imbalances present in the data. InfoNCE, despite preserving the highest latent structure, exhibits the lowest AUROC and Kappa, as anticipated due to its unsupervised formulation. SupCon inherits the latent structure preservation from InfoNCE while maintaining strong classification performance. SupCon matches the performance of black box algorithms, as evidenced by comparable AUROC (0.980 vs. 0.960) and Kappa scores (0.910 vs. 0.868).

Table 1. Classification Performance using EN-B2 Image Encoder and the best performing K (K = 25). BB = Black box algorithm.

Loss Functions	Ensemble AUROC				Ensemble Kappa			
	K = 25	NIR	I-NIR	BB	K = 25	NIR	I-NIR	BB
CE	0.978	**0.980**	**0.980**	0.960	0.880	0.870	0.880	0.868
InfoNCE	0.939	0.933	0.936	–	0.710	0.730	0.710	–
SupCon	0.969	**0.980**	**0.980**	–	0.910	**0.900**	**0.910**	–
Triplet	0.965	0.968	0.968	–	0.860	0.860	0.860	–
NCA	0.972	0.968	0.969	–	0.880	0.880	0.880	–

8.3 LLM Results

We perform the following experiments using only models trained with SupCon, due to the success of that training objective on retrieval and classification.

Summarization. While GPT-4 outperforms GPT-3.5 head-to-head in every respect, both LLMs produce more accurate, complete and relevant summaries for plain and normal reports (Fig. 4a). For sets of alternative reports, both often fail to perform correct mathematical operations (mean, min, and max) for tumor sizes expressed in different units, and to group together synonyms for the same anatomical plane.

Q&A. We note that GPT-4 significantly excels in the Q&A task (Fig. 4b). Q1 and Q3 involve mathematical analysis, while Q2 and Q4 focus on semantic understanding. GPT-4 is adept at simple math tasks like finding maximum values (Q3), but it struggles with more complex tasks like identifying similar tumor sizes (Q1). Both LLMs perform poorly on alternative reports, suggesting that variations in structure, units, and language challenge the LLM's ability to provide accurate answers.

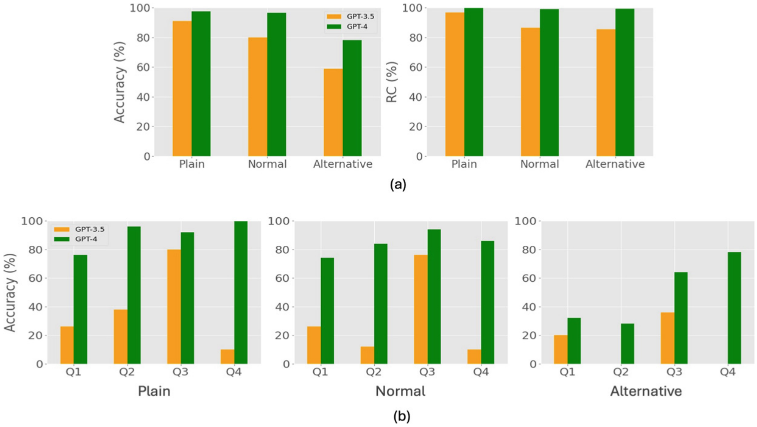


Fig. 4. LLM results. GPT-3.5 (Yellow) and GPT-4 (Green). (a) Summarization. The two subplots indicate the summarization performance using accuracy (left) and RC (right) (b) Q&A System. Each subplot in (b) indicates a type of report.

9 Discussion

In this paper, we catalog a search for the CBIR framework that maximizes interpretability while maintaining performance. For the models and objectives tests, we find that our best performing framework is an EN-B2 image encoder with interactivity enabled through GPT-4 which has SupCon training that leads to i) a structured retrieval space, and ii) maintains strong classification using NIR.

While the performance of SupCon is comparable to the black box model, its preservation of latent structure is ideal for facilitating usability and trust from clinicians. A key benefit of latent structure is that it develops automatically during training, eliminating the need for post-hoc labeling and finetuning. Future enhancements could involve clinicians ranking these retrievals to gauge their clinical relevance.

We feel that especially in the medical domain, it is important to build clinician-ML systems where models augment and support the decisions made by the doctor but are not the sole decision maker. Therefore, showing the predicted query class and confidence, the top retrievals, and summary of the retrievals on a platform where clinicians can interactively ask questions, could help those clinicians to better understand the model's reasoning (Appendix 7).

In our study, given that each patient has multiple scans, it is plausible to observe that smaller K retrievals may originate from a single patient. This situation introduces a potential bias, as the model's performance could be artificially inflated by the similarity within a single patient's scans rather than a genuine ability to generalize across diverse cases. To mitigate the influence of this bias, we extended our analysis by examining the classification performance at increased K values, ensuring our results reflect a more robust evaluation of the model's capabilities.

In our LLM experiments, we generated pseudo-text descriptions of scans due to dataset constraints, rather than using actual radiology reports. Despite this, our findings offer an initial insight into the promising integration of off-the-shelf LLMs with medical image analysis in radiology and anticipate that these trends will hold true with authentic radiology reports.

10 Conclusion and Future Work

We showcase a novel interactive CBIR system that accurately and interpretably classifies the type of brain tumor present in an MRI scan. Our usage of an off-the-shelf LLM broadens the potential for adoption of our system. While we document the clear limitations of these LLMs at present, the trend of improvement for these models inspires confidence in their future usability. Instead of using existing reports, some proposed systems [23] use multimodal LLMs to generate reports directly from medical images, but with this high level of complexity comes the risk of hallucination. We intentionally focus on leveraging an LLM strength reasoning over simple relationships–to augment interpretation of human-created content. Future work consists of applying this system to real-world users, full scan volumes, actual radiology reports, and to different modalities. This approach could be further extended in determining patient outcomes and treatment plans based on similar patients from the past, resulting in an improvement in patient care. Given that the enhanced CBIR system improves accessibility, this could open the door towards patients maintaining a more interactive role in their own care.

Appendix

1. Propensity of KNN and NIR to Overfit

Table 2 emphasizes KNN's increased propensity for overfitting the validation dataset when compared to NIR. In this experiment, a segment of the test dataset was allocated as a validation set. The KNN algorithm was applied to the validation set to determine the optimal value of K, which was subsequently utilized to compute Kappa for the test dataset. This procedure was replicated across five distinct, randomly selected validation subsets and the mean Kappa are reported. We see that the difference between Kappa for validation and testing datasets is notably greater for KNN as compared to NIR, supporting the evidence of KNN's greater tendency towards overfitting. Furthermore, NIR outperforms KNN in terms of Kappa on the test dataset.

Table 2. Assessing the propensity of KNN to overfit. This table shows the Cohen Kappa on the validation and testing.

KNN		NIR	
Validation	Testing	Validation	Testing
0.883	0.856	0.862	0.860

2. Textual Description Examples

Plain text: *The magnetic resonance imaging (MRI) T1-weighted scan in the axial. plane displays evidence of a meningioma with a size of 1097.02 mm^2.*

Normal report: *On axial T1 weighted MRI images, there is evidence of a meningioma with an area of approximately 1097.02 mm^2. The tumor characteristics are consistent with the typical appearance of a meningioma in this imaging modality.*

Alternate report: *On transverse T1 weighted MRI images, there is evidence of a meningeal tumor with an area of approximately 10.97 cm^2. The tumor characteristics are consistent with the typical appearance of a meningeal tumor in this imaging modality.*

3. LLM Prompts

3.1 Summarization

```
1)..Report 1 Content..
2)..Report 2 Content..
3)..Report 3 Content..
4)..Report 4 Content..
5)..Report 5 Content..
In these five reports, please identify the important at-
tributes that they have in common. Please write one sen-
tence about each of these attributes that summarizes/ag-
gregates the content in all reports. If any numbers are
involved then show the minimum, maximum and the average.
Imagine that someone doesn't have time to read all five
reports but wants to get a complete summarized version of
it. The important attributes should be, imaging modality,
imaging technique, tumor, tumor size, and plane.
```

3.2 Q&A

```
Query Report: ..Query Report..
Other Reports:
1)..Report 1 Content..
2)..Report 2 Content..
3)..Report 3 Content..
4)..Report 4 Content..
5)..Report 5 Content..
Question to ask
```

4. Metrics

4.1 Classification

Our classification objective is to assign a brain tumor type to each query image. We quantify performance using the area under the receiver operating characteristic (AUROC) and Cohen's Kappa (Kappa), a point accuracy metric adjusted for differences in class prevalence. Here, we calculate Kappa based on a soft ensemble for classification.

4.2 LLM Response Evaluation

In evaluating the tasks involving medical texts, we employ two metrics: Accuracy and relevance/completeness (RC). Accuracy assesses the correctness of the provided answers or information. It is essential in ensuring that the information conveyed is reliable and accurate, a non-negotiable requirement in medical contexts. RC evaluates whether all necessary attributes are appropriately mentioned and if the generated text pertinent to these attributes is relevant. This dual aspect of the metric not only considers the presence of key information but also scrutinizes its pertinence to the given context, thereby gauging whether a response is comprehensive and contextually appropriate. We evaluate the radiology report generation and summarization task using FA and RC and the Q&A system is using only FA. We score the Q&A system on 1 or 0—whether the LLM gets the answer right or wrong. Since the summarization consists of five attributes, we score each attribute 0 or 1. The one exception is for tumor size as we ask the LLM to provide the min, max, and the average tumor size across the five radiology reports. We weight the scores for them as 0.25, 0.25, and 0.5 respectively (total for tumor size adds up to 1). Each summary is graded on a total score of 5 for each metric.

5. Experiment Design

We perform two types of data augmentation during training, namely random Gaussian blur and random rotate, effectively increasing the size of the training data by three. The augmented images are not used for inference. As a data preprocessing step, we normalized each image between -1 and 1, resampled to a size of 1 x 224 x 224, and duplicated the first channel three times in order to support the use of pre-training weights on models originally trained on 3-channel images.

Code was implemented in Pytorch and used a NVIDIA RTX A6000 GPU for training and inference. We performed hyperparameter searches over learning rate, temperature, margin, and epoch and report the final performance of each model as the mean across 5 trials (i.e., training runs) to support reproducibility. The batch size set for training dataset was 30. For each dataset, 3-fold cross validation was done to obtain the optimal hyperparameters. Performance is reported on withheld test images not used for training or hyperparameter optimization.

6. Detailed Classification Results

We experimented with two convolutional neural network-based model architectures: Resnet50 and EF-B2 (Table 3 and Table 4). To train the image encoder, we explored classification loss (cross entropy (CE)) and metric learning losses such as triplet loss, neighborhood component analysis (NCA), InfoNCE (INCE) and supervised contrastive loss (SupCon).

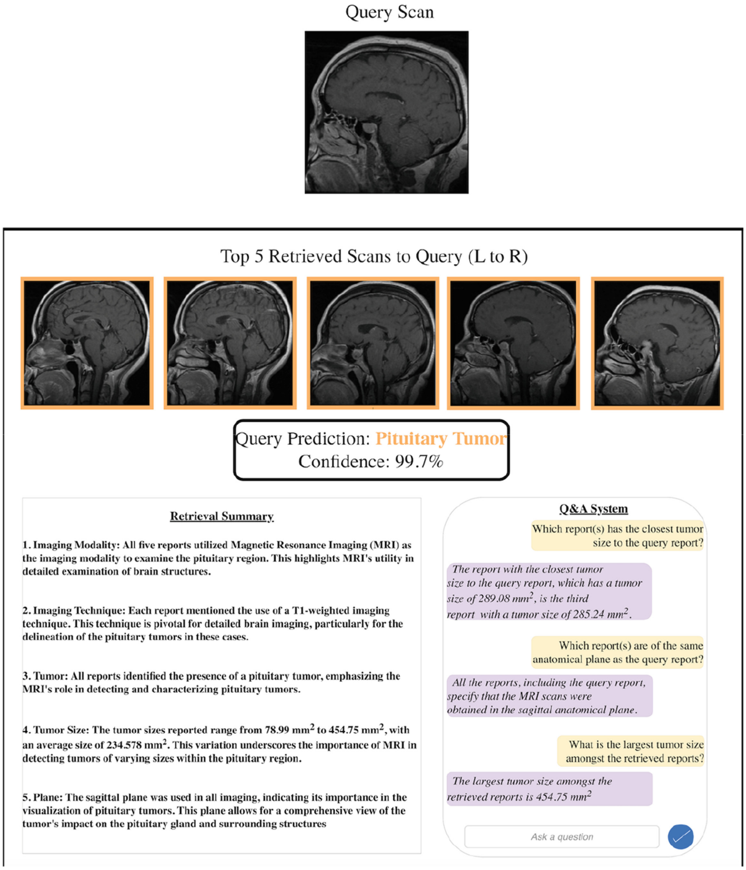
Table 3. Cohen Kappa scores across model architectures. A-CK is the average Kappa while E-CK is the ensemble kappa stat across 5 trials.

		EF-B2		ResNet50	
		A-CK	E-CK	A-CK	E-CK
K = 5	CE	0.825	0.863	0.761	0.818
	NCA	0.808	0.854	0.758	0.822
	Triplet	0.808	0.868	0.792	0.837
	SupCon	0.893	**0.906**	0.799	**0.843**
	INCE	0.693	0.725	0.543	0.616
K = 25	CE	0.834	0.878	0.768	0.816
	NCA	0.818	0.858	0.765	0.832
	Triplet	0.815	0.868	0.792	0.838
	SupCon	0.894	**0.907**	0.798	**0.838**
	INCE	0.702	0.735	0.509	0.559
MLP		0.781	0.868	0.729	0.794

7. Output Example

Table 4. AUROC scores for EF-B2 model architecture across 5 different training loss functions.

Loss Function	AUROC					
	K = 5	K = 7	K = 25	NIR	I-NIR	MLP
CE	0.971	0.975	0.978	**0.98**	**0.98**	0.96
Triplet	0.964	0.964	0.965	0.968	0.968	–
NCA	0.967	0.968	0.972	0.968	0.969	–
InfoNCE	0.932	0.934	0.939	0.933	0.936	–
SupCon	0.969	0.969	0.969	**0.98**	**0.98**	–


Query Scan
Top 5 Retrieved Scans to Query (L to R)
Query Prediction: Pituitary Tumor
Confidence: 99.7%
Ask a question

Retrieval Summary

1. Imaging Modality: All five reports utilized Magnetic Resonance Imaging (MRI) as the imaging modality to examine the pituitary region. This highlights MRI's utility in detailed examination of brain structures.

2. Imaging Technique: Each report mentioned the use of a T1-weighted imaging technique. This technique is pivotal for detailed brain imaging, particularly for the delineation of the pituitary tumors in these cases.

3. Tumor: All reports identified the presence of a pituitary tumor, emphasizing the MRI's role in detecting and characterizing pituitary tumors.

4. Tumor Size: The tumor sizes reported range from 78.99 mm^2 to 454.75 mm^2, with an average size of 234.578 mm^2. This variation underscores the importance of MRI in detecting tumors of varying sizes within the pituitary region.

5. Plane: The sagittal plane was used in all imaging, indicating its importance in the visualization of pituitary tumors. This plane allows for a comprehensive view of the tumor's impact on the pituitary gland and surrounding structures

Q&A System

Which report(s) has the closest tumor size to the query report?

The report with the closest tumor size to the query report, which has a tumor size of 289.08 mm^2, is the third report with a tumor size of 285.24 mm^2.

Which report(s) are of the same anatomical plane as the query report?

All the reports, including the query report, specify that the MRI scans were obtained in the sagittal anatomical plane.

What is the largest tumor size amongst the retrieved reports?

The largest tumor size amongst the retrieved reports is 454.75 mm^2

Fig. 5. Overview of the entire system: Retrieval, Classification, and LLM based Summary and Q&A

References

1. DeAngelis, L.M.: Brain tumors. N. Engl. J. Med. **344**, 114–123 (2001). https://doi.org/10.1056/NEJM200101113440207
2. Bauer, S., Wiest, R., Nolte, L.-P., Reyes, M.: A survey of MRI-based medical image analysis for brain tumor studies. Phys. Med. Biol. **58**, R97–R129 (2013). https://doi.org/10.1088/00319155/58/13/R97
3. Goldfarb, A., Teodoridis, F.: Why is AI adoption in health care lagging? Brookings Institution (2022)
4. Hosny, A., Parmar, C., Coroller, T.P., et al.: Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLOS Med **15**, e1002711 (2018). https://doi.org/10.1371/journal.pmed.1002711
5. Lanjewar, M.G., Parab, J.S., Shaikh, A.Y.: Development of framework by combining CNN with KNN to detect Alzheimer's disease using MRI images. Multimed Tools Appl **82**, 12699–12717 (2023). https://doi.org/10.1007/s11042-022-13935-4
6. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017). https://doi.org/10.1038/nature21056
7. Sendak, M.P., Ratliff, W., Sarro, D., et al.: Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med. Inform. **8**, e15182 (2020). https://doi.org/10.2196/15182
8. Lu, S.-C., Swisher, C.L., Chung, C., et al.: On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Front. Oncol. **13**, 1129380 (2023). https://doi.org/10.3389/fonc.2023.1129380
9. Choe, J., Choi, H.Y., Lee, S.M., et al.: Evaluation of retrieval accuracy and visual similarity in content-based image retrieval of chest CT for obstructive lung disease. Sci. Rep. **14**, 4587 (2024). https://doi.org/10.1038/s41598-024-54954-5
10. Herrmann, A.E., Estrela, V.V.: Content based image retrieval (CBIR) in remote clinical diagnosis and healthcare (2016)
11. Hu, B., Vasu, B., Hoogs, A.: X-MIR: EXplainable Medical Image Retrieval. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1544–1554. IEEE, Waikoloa, HI, USA (2022)
12. Qayyum, A., Anwar, S.M., Awais, M., Majid, M.: Medical image retrieval using deep convolutional neural network. Neurocomputing **266**, 8–20 (2017). https://doi.org/10.1016/j.neucom.2017.05.025
13. Owais, M., Arsalan, M., Choi, J., Park, K.R.: Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence. J. Clin. Med. **8**, 462 (2019). https://doi.org/10.3390/jcm8040462
14. Codella, N.C.F., et al.: Collaborative human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images (2018) https://doi.org/10.1007/978-3-030-02628-8_11
15. Arun, N., Gaw, N., Singh, P., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell **3**, e200267 (2021). https://doi.org/10.1148/ryai.2021200267
16. Cheng, J., Huang, W., Cao, S., et al.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE **10**, e0140381 (2015). https://doi.org/10.1371/journal.pone.0140381
17. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)

18. Goldberger, J., Hinton, G.E., Roweis, S., Salakhutdinov, R.R.: Neighbourhood Components Analysis. In: Advances in Neural Information Processing Systems. MIT Press (2004)
19. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019)
20. Khosla, P., et al.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 18661–18673 (2020)
21. Zulfiqar, F., Ijaz Bajwa, U., Mehmood, Y.: Multi-class classification of brain tumor types from MR images using EfficientNets. Biomed. Signal Process. Control **84**, 104777 (2023). https://doi.org/10.1016/j.bspc.2023.104777
22. Ramaswamy, V.V., Kim, S.S.Y., Fong, R., Russakovsky, O.: Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10932–10941. IEEE, Vancouver, BC, Canada (2023)
23. Lu, Y., Hong, S., Shah, Y., Xu, P.: Effectively fine-tune to improve large multimodal models for radiology report generation (2023)