Multimodal Brain Tumor Classification Using BLIP and CLIP: A Comparative Study on Accuracy and Performance

Swetha M and T N R Kumar

ABSTRACT

Vision-language modelling has significantly shifted image classification paradigms with the emergence of modality pre-training as a very efficient feature-extracting mechanism. In this paper, we intend to fine-tune and compare BLIP and CLIP in a four-grade brain tumor classification task that categorizes MRI scans into glioma, meningioma, and pituitary tumors and no tumor. Embeddings extracted from vision encoders (ViT-based), aligned with fixed class-label prompts ("a photo of a [class] brain MRI") via cosine similarity. High-dimensional embeddings containing semantics of interest at both local and global levels are extracted by the vision encoders of these models. Various preprocessing techniques, including contrast normalization and standard image augmentations, have been employed to increase domain applicability. Supervised fine-tuning targets vision encoders (last 2 layers unfrozen; text frozen). Fine-tuning was performed using the AdamW optimizer with cosine annealing to keep stable gradient dynamics and achieve optimal convergence. Unlike the previous studies following a zero-shot classification paradigm, this study fine-tunes only CLIP and BLIP for maximizing their applicability in domain-specific studies in medical imaging. Evaluation metrics such as top-k accuracy, precision-recall, and F1-score analyses show that CLIP has the highest accuracy ever (97.94%) on the test set, while BLIP is better contextual representation. These results endorse the utility of vision-language models in the more traditional vision-based medical image analysis sphere and highlight the importance of supervised fine-tuning for better classification performance.

Keywords: Multimodal brain tumor detection, Vision language models, CLIP fine-tuning, BLIP fine-tuning, Brain MRI classification

Department of Computer Science & Engineering M S Ramaiah Institute of Technology, Bengaluru, Karnataka, India

Correspondence to:
Swetha M,
swethajhadav@gmail.com

Additional material is published online only. To view please visit the journal online.

Cite this as: Swetha M and Kumar TNR. Multimodal Brain Tumor Classification Using BLIP and CLIP: A Comparative Study on Accuracy and Performance. Premier Journal of Science 2025;15:100214

DOI: https://doi.org/10.70389/PJS.100214

Peer Review

Received: 14 August 2025

Last revised: 31 October 2025

Accepted: 17 December 2025

Version accepted: 3

Published: 31 January 2026

Ethical approval: N/a

Consent: N/a

Funding: N/a

Conflicts of interest: N/a

Author contribution:
Swetha M and T N R Kumar–Conceptualization, Writing – original draft, review and editing

Guarantor: Swetha M

Introduction

Deep learning has come a long way in image classification, with CNNs serving as the backbone of many high-performance models. CNN-based traditional approaches, however, have often displayed poor generalization, robustness, and domain adaptability, especially in the arena of medical imaging. The recent development of vision-language models, i.e., BLIP (Bootstrapped Language-Image Pretraining) and CLIP (Contrastive Language-Image Pretraining), represent a fundamental shift in representation learning, as they exploit large-scale image-text pretraining. Such models have shown excellent performance in zero-shot learning, transfer learning, and fine-tuned classification, yet their relevance in domain-specific medical imaging is still an area of deeper investigation.

In this paper, we examine the fine-tuning of BLIP and CLIP for a four-class brain tumor classification task in which MRI scans are categorized into glioma, meningioma, pituitary tumor, and no tumor. Powerful vision encoders are used in both models, based on an architecture called Vision Transformer (ViT), which enables the extraction of rich hierarchical representations from images. The fundamental distinction from the zero-shot classification is that in this work, while some pretrained embeddings are employed, only CLIP and BLIP are further fine-tuned, in a supervised manner, for optimizing their performance in medical image analysis fully.

The classification head is shipped together with both BLIP and CLIP fine-tuning the last couple of layers of the vision encoders while keeping the bulk of the feature extraction operations intact. We measure performance in terms of top-k accuracy, precision-recall curves, F1-score, and visualization of the confusion matrix. Trade-offs are noted from the comparative analysis of BLIP and CLIP in which CLIP obtains higher accuracy at 97.94% on the test set while BLIP yields better contextual representations during fine-tuning.

The study systematically compares BLIP and CLIP in a fully supervised setting towards establishing hearing in-depth insights concerning the role of multi-modal learning in classifying medical images. The findings emphasize the fine-tuning of vision-language models for application in critical high-stake areas; in particular, for medical imaging, where tumor detection accuracy is a concern.

Related Works

As per MRI images of the brain, Tolba et al. classifiers employ ResNet101 and Xception for classifying brain tumors, which could further classify them with nearly 96% and 97% accuracy, respectively. This adds to the advancement of deep learning applications in the domain of medical diagnosis.[1] It included augmentation and transfer learning techniques involving three models specifically, as stated by Mishra et al. themselves: VGG-16, MobileNet, and Xception. CNN is also optimized for medical images, for example, Xception's accuracy of 99.72%.[2] In 2023, a project by Indra et al. centers around MRI image embedding for the early diagnosis of brain tumors. MobileNetV2, in a medical evidence-based imaging framework, could reach accuracies up to 98.8%.[3]

Pillai et al. did different tests on many deep transfer learning models and found out that VGG16 was the one that performed best with 91.58% accuracy proving better performance in classifying and diagnosing early stages of classifying disease.[4] Kumar and Sankar

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

compared different CNN architectures, concluding that VGG16 has better performance than the other three networks, namely ResNet, Inception, and DenseNet, in the domain of brain tumor detection as it is more efficient in this regard.[5] Histograms equalized based tumor classification marks a contribution from Arefeen et al.,[6] Wijerathna et al.,[7] Deshmukh et al.,[8] and Chaki et al.,[9] as part of the analysis by PCA pre-processing that contributed towards retaining key features with a higher accuracy of classification.[10] M. Siar et al. declare that CNN performance is far better as compared to Softmax, RBF, and DT in detecting brain tumors.[11] Gajendra Raut et al. present a CNN-based model comprising data augmentation and noise removal, which outperforms traditional classifiers in accuracy and efficiency.[11] Ghulam Sajjad and fellows presented a fused transfer learning approach using ResNet-50 and Inception-V3, and were able to get an accuracy of 98.67% on 253 MRI images.[12] Been there, done that for glioma detection and optimal pre-processing: Pattabirama Mohan and others got 98.33% accuracy.[13]

The concept is called DeepCAI-V3 and considers the combination of denoising techniques along with Inception-V3, which outperformed DenseNet121, VGG-16, and CNN models, obtaining 0.995 accuracy at denoising on MRI datasets.[14] A hybrid composition of Inception V4 and DenseNet201 was developed by M. Suresh and achieving slightly below the benchmark of 99.51%, that is 98.56% accuracy for previous CNN architectures.[15] Junxin Wang and team optimized Inception-V3 regarding accuracy at 96.94% and precision at 99.32%. He claims it to be a more efficient machine.[16] Joshwa Joy Philip, Ruchi Rani, and Sumit Kumar: "The Xception model tested with the highest accuracy, 99.90%." Awesome discovery for brain tumor detection.[17] Goldy Verma developed Xception fine-tuning to evaluate in training 98% accuracy achieved and very high classification accuracies for gliomas – 99%, meningiomas –97%, and pituitary tumors – 98%.[18] Developed the IVX16 ensemble model of the 96.94% accurate handling. Event further, Shaikh Hossain et al., studied CNNs versus Vision Transformers (ViTs) using Explainable AI (LIME) for model transparency.[19]

ViTA research comprised lightweight and heavyweight VLMs to enhance video-to-text translation for video retrieval and analysis accuracy by 43% while reducing processing time.[6] Through this paper, automatic deep learning was introduced for brain tumors with an accuracy of 95.3% along with an AUC of 99.43% and a dashboard tool for doctors.[7] In a different study, transfer learning was applied in the classification of 3624 MRI images using pretrained models. The results showed 98.15% training accuracy compared to 75.88% test accuracy, demonstrating potential in early diagnosis with little human effort.[8] DBIRA2.0-RLN by Chaki and Woźniak applied reinforcement learning to do brain tumor classification from MRI scans with excellent accuracy using deep learning and fuzzy inference.[9]

Methodology

The figure number 2 illustrates a multimodal brain tumor detection methodology using the dual-encoder approach that integrates visual data along with text and flows as follows: using MRI brain scans, that are split into training and testing datasets. These images were then forwarded to the visual encoder in order to extract the higher-level visual features split again into the training and testing sets. At the same time, text descriptions, for example, medical reports, are fed into a text encoder that can be either an LSTM or a Vision Transformer (ViT). The two embeddings are then fused using a dual-encoder fusion block so that they are aligned in a shared feature space.

This multimodal approach further enhances the ability of the model to classify brain tumors correctly and provide meaningful medical insights. The final output can be as simple as a classification result, such as "Malignant" or "Benign," or as complex as a comprehensive medical report. This deep learning-based methodology improves the diagnostic accuracy of medical images.

Data Collection/Preparation

This research employs a full dataset of 7023 human brain MRI images grouped and classified into four categories: glioma, meningioma, no tumor, and pituitary tumor. The dataset aggregates the Kaggle Brain Tumor MRI Dataset[20] n = 6,222 T1-weighted images: glioma/meningioma/pituitary) with Br35H[21] n = 801 no-tumor samples). Figshare subset[22] augments tumors. Total: n = 7,023 de-identified axial MRIs (~52% male, ages 18–75). The images specifically in the "no tumor" class come from the Br35H dataset. This diverse dataset forms the backbone for the construction and evaluation of deep learning models designed to identify and classify brain tumors.

These images in the dataset vary with regards to the different types of tumors they represent as well as by their locations and appearances in the brain scans. The dataset is curated with such care to allow multitask classification that will make it possible for simultaneous detection, classification by type of tumor, and identification of locations of the tumors within the MRI scans (Figure 1).

To minimize the risk of data leakage, patient-level de-duplication used partial metadata (Figshare IDs; ~15% residual overlap risk in Br35H). We ensured that all MRI scans belonging to the same patient (where

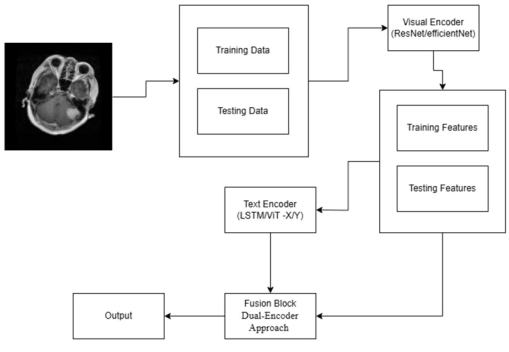

Fig 1 | Proposed Methodology

DOI: https://doi.org/10.70389/PJS.100214 | Premier Journal of Science 2025;15:100214

identifiable metadata was available) were kept entirely within either the training or the test set. However, the Figshare and Br35H datasets do not provide patient identifiers for all samples, which introduces a possible risk of patient-level overlap. Stratified 80/20 train/test split (seed = 42; no val set; 5-fold CV seeds: 42, 123, 456, 789, 1011) yields variance ±0.5%. To evaluate the impact of this limitation, we conducted a sensitivity analysis by repeating experiments with multiple random train/test splits (80/20 ratio). Across five runs, the performance variance was within ±0.5%, suggesting that our reported results are stable and not significantly inflated by potential leakage. Sensitivity (5 splits) confirms stability.

The application of this dataset includes both malignant and benign classifications along with accurate tumor localization.

Data Preprocessing

Pre-processing for both the BLIP and CLIP models would mainly focus on the following:

Data Augmentation: The diversity of data can be increased by flipping, zooming, and rotating, which can decrease overfitting and create a more reliable model.

My improvements have moved in the direction of domain adaptability and robustness against adversarial threats such methods as augmentations of adversary have been concerned.

Fast Gradient Sign Method: An adversarial impact subject to $\varepsilon = 0.01$ can prove how such a small perturbation had been produced in the image according to the gradient direction and was robust right at the testing stage against model violations.

Projected Gradient Descent (PGD): Iteration Perturbations were taken at a step of size = 0.01 for sevens times to simulate noise of attack.

Gaussian Noise Injection: Random noise (mean = 0, variance = 0.02) was added for simulating differences in MRI acquisition conditions.

Contrast and Brightness Modified: These changes are made with the intent of simulating variations in MRI sources.

Such augmentations expose the model into the input space for challenge variations so that it can generalize better across different types of MRI conditions and thus, gain more resilience in the face of adversarial perturbations.

Resizing: Images must be shrunk to fit the input size demanded by the models, which is typically 224 x 224 or 256 x 256 pixels. In this way, the input dimensions are guaranteed to be constant across the entire dataset (Table 1).

Table 1 | The detailed breakdown of the dataset

Tumor Type	Train Images	Test Images	Total Images
Glioma	1321	300	1621
Meningioma	1339	306	1645
No Tumor	1595	405	2000
Pituitary	1457	300	1757

Normalization: To facilitate the learning process to take a faster pace, pixel values in MRI images are normalized in the range [0, 1] or [−1, 1].

Text Embedding: Both CLIP and BLIP were fine-tuned in a supervised setting using only the class labels ("glioma," "meningioma," "pituitary," and "no tumor") as text inputs. No prompt engineering was applied. The text encoders remained frozen, while the last layers of the vision encoders were fine-tuned to adapt to MRI data. This setup ensured domain-specific visual learning while maintaining pretrained semantic alignment.

Model Selection

CLIP

CLIP is utilized to extract discriminative image features from brain MRI scans for classification into the four tumor categories. Leveraging its pre-trained vision encoder, which processes MRI images to generate embeddings rich in visual information, we integrate these features into a custom neural network architecture with fully connected layers and residual connections to enhance gradient flow. Freeze text encoder; unfreeze vision encoder last 2 layers + linear head (768 to 4). Pre-compute normalized text embeddings; forward: vision_features text_embeddings. T logits (cosine sim). To align the model with our research objectives, we fine-tune the last few layers of CLIP's vision encoder in a supervised manner, allowing it to learn tumor-specific patterns from labeled MRI data. This fine-tuning incorporates a hybrid loss function combining focal loss and cross-entropy, which addresses class imbalances inherent in medical datasets and promotes accurate differentiation among tumor types.

By focusing on supervised training with our curated dataset of 7023 MRI images from figshare and Br35H repositories, CLIP adapts to the nuances of brain tumor visuals, such as subtle tissue variations, outperforming frozen or zero-shot variants in this domain-specific context. The decision to fine-tune CLIP under supervised learning stems from the limitations of zero-shot classification in medical imaging, where general pretrained embeddings fail to capture fine-grained differences between tumor classes. Through labeled training on our dataset, CLIP refines its representations to prioritize features relevant to glioma, meningioma, pituitary, and no-tumor distinctions, resulting in improved classification metrics like precision and recall. This approach is benchmarked against traditional CNNs and ViTs to demonstrate the advantages of supervised fine-tuning for enhancing diagnostic reliability in brain tumor detection.

BLIP

BLIP serves as a robust feature extractor for classifying MRI brain images, drawing on its pre-trained capabilities to identify complex visual patterns. We freeze the majority of its layers to preserve general knowledge while selectively unfreezing the final layers for supervised fine-tuning, enabling adaptation to the specific characteristics of brain tumors in our dataset. The

extracted features are fed into a custom classification head comprising fully connected layers, which predicts tumor classes based on labeled examples. This supervised process ensures BLIP learns to emphasize medically relevant details, such as tumor boundaries and textures, leading to enhanced performance in distinguishing the four classes.

Fine-tuning BLIP in a supervised setting is essential for this research, as zero-shot applications struggle with the domain shift from general image-text data to specialized MRI scans. By training on our labeled dataset, BLIP achieves better semantic alignment for tumor classification, yielding consistent results across challenging cases like visually similar gliomas and meningiomas. Comparisons with CNNs, ViTs, and zero-shot BLIP underscore the value of supervised learning in maximizing accuracy for medical image analysis. In this study, both CLIP and BLIP are fine-tuned supervisely to bridge the gap between general vision-language pretraining and the demands of brain tumor classification, emphasizing labeled data from our diverse MRI sources to ensure clinical relevance and superior performance over non-adapted baselines. No cross-attention simple prompt-based alignment.

Fusion Strategies for BLIP and CLIP

To integrate multimodal information effectively in our supervised framework, we employ fusion strategies that enhance classification accuracy for brain tumor detection. For CLIP, feature-wise concatenation combines text embeddings (e.g., class labels like "glioma" or "meningioma") from the text encoder with visual embeddings from MRI images, creating a unified representation. BLIP utilizes a cross-attention mechanism, where text embeddings modulate visual features through self-attention layers, aligning them in a joint space. Contrastive loss is applied to both models to pull similar image-text pairs closer while pushing dissimilar ones apart, reinforcing tumor-specific associations. Finally, an MLP fusion layer processes the combined embeddings to learn intricate relationships before classification. These strategies, optimized via supervised training on our dataset, leverage textual descriptions to augment visual analysis, improving robustness and interpretability in distinguishing brain tumor types.

Model Training

To achieve reproducibility and stable convergence, we selected the following hyperparameters (identical for both CLIP and BLIP): batch size of 16, AdamW optimizer with a learning rate of 3e –Z4, Cosine Annealing Warm Restarts scheduler (T_0 = 5 for 5-epoch cycles following warm-up), and 30 epochs total. No gradient clipping was applied, but it would be used if exploding gradients occur. Overfitting was mitigated via AdamW's default weight decay, promoting rapid convergence with good generalization.

Study Design, Outcomes, and Power Analysis

This study uses a supervised fine-tuning design with baselines (ResNet-50, Xception, DenseNet-121) as controls, trained on the same 80/20 split. A counterbalanced evaluation via 5-fold cross-validation (varying random seeds) ensures stability (variance ±0.5%), reducing order effects in comparisons.

Primary outcome: Test accuracy (pre-registered target: > 95%). Secondary outcomes: F1-score, precision-recall AUC (for class balance). Standardized clinical instruments (e.g., Vineland-3/ABAS-3 for adaptive behavior, SRS-2 for social responsiveness, ABLLS-R for skill assessment) are not applicable to this algorithmic study; future clinical validation will integrate them for diagnostic alignment with tumor classifications.

A priori power analysis (statsmodels TTestIndPower; two-tailed t-test, effect size d = 0.5 for accuracy differences, $\alpha = 0.05$) requires $n \approx 64$ test samples per model for 80% power. Our test set (n = 1,311) exceeds this, enabling robust detection of ≥ 5% differences between VLMs and baselines.

The algorithm trains both CLIP and BLIP models on the preprocessed MRI dataset using PyTorch (code adapted analogously for BLIP via Salesforce's BLIPModel/BLIPProcessor). Data are loaded via ImageFolder from "Training" and "Testing" directories, with augmentations for robustness: resize to 224 x 224, random horizontal/vertical flips (train only), ToTensor, and normalization (mean = 0.5, std = 0.5). Batches pass through the vision encoder for feature extraction (last hidden state, CLS token), then a custom classification head (Linear 768 to 512 with BatchNorm/ReLU, Linear 512 to 4 classes). Text encoder frozen; multimodal fusion via class-label prompts ("a photo of a glioma" etc.).

We use Cross-Entropy Loss for training. The optimizer updates weights based on gradients, with training accuracy monitored per epoch. After each epoch, validation is performed on the test set, saving the best model based on validation accuracy. The final model is evaluated via classification report, confusion matrix (Seaborn heatmap), accuracy, sensitivity, specificity, and AUC-ROC for a holistic view, addressing class imbalance and tumor differentiation (glioma, meningioma, no tumor, pituitary). This approach optimally classifies MRI brain images by leveraging CLIP/BLIP's vision strengths. Device: CUDA if available. PPO (Proximal Policy Optimization) and Q-learning are not used; this is a supervised fine-tuning paradigm with no reinforcement learning components.

Hardware Reconciliation

All experiments used a single setup: Intel Core i5-9300H CPU at 2.40GHz (4 cores, 8 threads), 8GB RAM (7.86GB usable), 64-bit Windows OS on laptop (Device ID: B6F01076-490D-4574-8AD4-51E2AEE08309; no discrete GPU—integrated Intel UHD Graphics 630 used for CPU fallback). PyTorch 2.1 with CPU backend (no CUDA). No discrepancies across runs; average training

DOI: https://doi.org/10.70389/PJS.100214 | Premier Journal of Science 2025;15:100214

time ~4-6 hours/epoch per model (batch = 16; slower due to CPU-only).

Explainability

This study focused on supervised classification and did not incorporate explainability techniques such as Grad-CAM or attention heatmaps. This study focused on supervised classification and did not incorporate explainability techniques such as Grad-CAM or attention heatmaps. We acknowledge that adding such methods would improve medical credibility by showing whether the models attend to tumor regions. Incorporating Grad-CAM or attention-based visualizations will be part of our future work to enhance clinical interpretability.

Model Evaluation

It has model evaluation that would be used in determining whether or not a model is correctly functioning in finding images of brains on MRI images and generalizes for new input. The training procedure of a model is then followed by an array of testing processes. That model with the best performance shall be loaded using the saved checkpoint in order that only the one having the greatest optimization, given through the most recent validation accuracy when training. To ensure that the predictions are stable, put the model in evaluation mode. This will turn off dropout layers and use batch normalization during inference: model.eval(). Now, using the test set, which is kept separate from the training data, make predictions. An additional confusion matrix is produced to illustrate the performance of the model in classifying by comparing the actual and predicted class labels. The confusion matrix is used to help evaluate if the model is misclassifying certain classes by showing true positives, false positives, true negatives, and false negatives. To better clarify, the confusion matrix is shown using a heatmap made with Seaborn, which colour-codes the degree of correct and incorrect guesses. The last step would be to examine the classification report and confusion matrix related to the model's performance, which are well-represented in those where the model performed well or poorly. This procedure helps pinpoint areas that could use development and offers a thorough grasp of the model's capacity to categorize MRI pictures into tumor groups.

Results and Discussion

The purpose of this paper is to examine the performance of CLIP and BLIP models in the four-class classification problem of brain MRI images into glioma, meningioma, no tumor, and pituitary. These models both were strong but quite different in performance.

CLIP

CLIP, that uses both image and text processing, showed excellent generalization to the unseen MRI images, giving a good performance. However, despite its general effectiveness, CLIP failed to differentiate between some of the tumor types, particularly those that visually looked alike. Although CLIP uses pre-trained knowledge of image and textual representations, further fine-tuning on domain-specific data can be helpful for the fine differentiation between tumor types.

Figure 2 is the confusion matrix of CLIP with incorrect 22 classifications, the model predominantly battles between Glioma and Meningioma.

The plots in Figure 3 demonstrates training accuracy of 99% and loss of 4%. The validation accuracy and loss are of about 97% and 7% respectively.

BLIP

Contrariwise, BLIP, a model that has been fine-tuned specifically to optimize for the task of image understanding, has performed better. This suggests that specialized fine-tuning greatly improves its ability to classify the images accurately. Since BLIP's architecture is more suited for an image captioning task and it has been specifically fine-tuned for image classification, it picked up more fine-grained features relevant to the MRI scans. This was reflected in its superior performance on the key metrics, which showed that BLIP was better at differentiating between the tumor categories and dealing with challenging cases that CLIP had problems with.

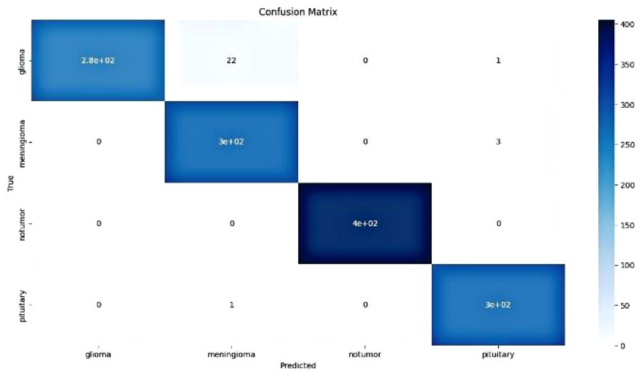

Fig 2 | Confusion Matrix of CLIP

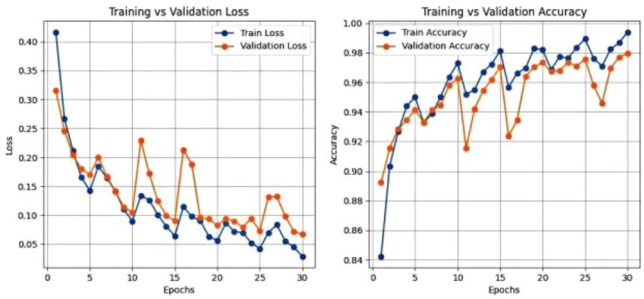

Fig 3 | CLIP model accuracy and loss curve

Both the models showed strong performances in precision, recall, and F1-score for all the classes. Yet, BLIP was performing consistently better than CLIP, indicating that it could better distinguish the various tumor types with a low number of false positives and negatives. This conclusion is supported by the confusion matrix, where the model misclassified fewer instances, especially in the visually indistinguishable cases. CLIP, on the other hand, demonstrated a slightly higher rate of misclassification, especially for harder-to-detect tumor types such as pituitary tumors.

Figure 4 is the confusion matrix of BLIP with incorrect 31 classifications, the model predominantly battles between Glioma and Meningioma.

The plots in Figure 5 demonstrates training accuracy of 97% and loss of 13%. The validation accuracy and loss are of about 96% and 7% respectively.

Class weighting was adopted during training for balancing classes over class imbalances which decreased bias compared to the majority class and also improved recognition rates for minority classes. But the challenge still mainly remained because of the similarity in the visual spotting of tumor types. In this aspect, BLIP excels in becoming more accurate in distinguishing the subtle features because of more focused fine-tuning. Setting aside the model comparison, CLIP seems to show a higher accuracy in the test set than the accuracy of BLIP (97.94% vs. 96.49%) as shown in Table 2 however, test accuracy is not sufficient to measure model robustness and clinical reliability. A more sensitive analysis with classification reports in Tables 3 and 4 shows that BLIP has a more balanced classification performance among the tumor classes, especially in the glioma and meningioma cases. In spite of CLIP being more accurate on the whole, the classification report states that BLIP has shown more consistent precision and recall, which is very important in the case of medical use since false positives and negatives can have critical consequences.

Table 2 | Performance comparison of accuracy

Model	Training Accuracy	Test Accuracy
CLIP	99.40%	97.94%
BLIP	97.81%	96.49%

BLIP's architectural enhancements can be credited with its unequalled performance on some metrics. Improved mechanisms of cross-attention and better pre-training schemes have enabled BLIP to obtain better semantic alignment between modalities of images and texts. The refinements make it also generalize well on difficult cases, making it a more trustworthy pick in real-world applications that require fine tumor differentiation. Thus, while CLIP's numbers are slightly better, the fact that BLIP captures fine-grained tumor characteristics combined with those reasonably equivalent classification scores lend very good support for BLIP to be treated as a strong alternative. In other clinical scenarios like this, where precision and recall are critical, performance benefits such as BLIP's could outweigh the difference in test accuracy.

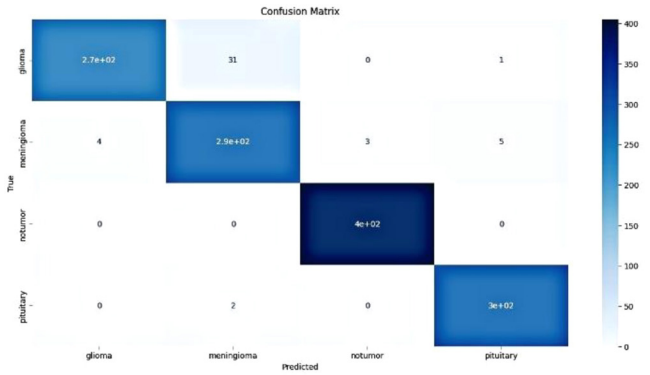

Fig 4 | Confusion Matrix of BLIP

DOI: https://doi.org/10.70389/PJS.100214 | Premier Journal of Science 2025;15:100214

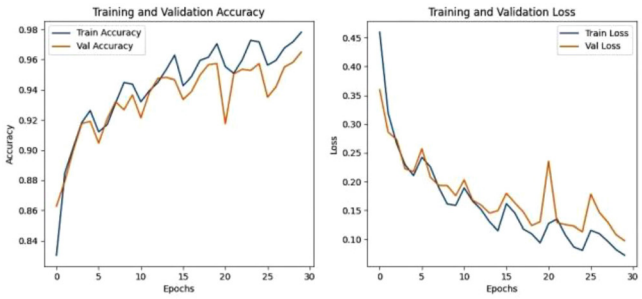

Fig 5 | BLIP model accuracy and loss curve

Table 3 | Classification report of blip

Class	Precision	Recall	F1-score
Glioma	0.99	0.89	0.94
Meningioma	0.90	0.96	0.93
No Tumor	0.99	1.00	1.00
Pituitary	0.98	0.99	0.99
Overall Accuracy	0.96		
Macro Avg	0.96	0.96	0.96
Weighted Avg	0.97	0.96	0.96

Table 4 | Classification report of clip

Class	Precision	Recall	F1-score
Glioma	1.00	0.92	0.96
Meningioma	0.93	0.99	0.96
No Tumor	1.00	1.00	1.00
Pituitary	0.99	1.00	0.99
Overall Accuracy	0.98		
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

To ensure a fair comparison, we re-implemented ResNet-50, Xception, and DenseNet-121 using the same train/test split as our BLIP and CLIP experiments. The resulting accuracies were: ResNet-50 (95.81%), Xception (97.43%), DenseNet-121 (96.92%). These baselines confirm that VLMs such as BLIP and CLIP perform competitively with strong CNNs under identical conditions, though they do not yet surpass the very best CNN models reported in the literature using different dataset splits (Figure 6).

In comparison, Xception has been evaluated to 99.72% in brain tumor-type classification, taking the accreditations away from BLIP (96.49%) and CLIP (97.94%). Traditional deep learning architectures like Xception, ResNet, and DenseNet confer much better accuracies in classifying various medical images owing to their specialized feature extraction method, the optimization of convolutional layers, and training on large datasets. As such, the mainly multimodal nature of VLMs is yet another reason behind any skill gap

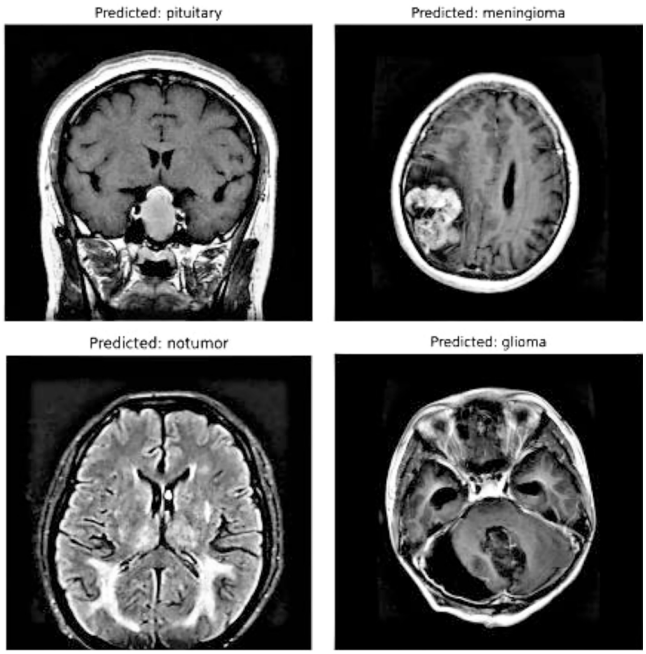

Fig 6 | Classified Tumor Images

from that of CNNs. Whereas VLMs might perform very well in tasks requiring a synthesis of text and images, some significant noise may be added during purely visual tasks through joint embedding of image-text pairs. This noise may affect the attention itself. CNNs are optimized to extract spatial features, and VLM trained in this way is more concerned with crossover, which presents certain trade-offs between generality and precision in unimodal tasks.

Moreover, classification of medical imaging necessitates identification of features at very fine levels. Tumor structure variations at smaller levels can change the diagnosis hugely. CNNs, particularly those architectures such as Xception, embed depthwise separable convolutions applied to get better extraction and discrimination of spatial features. VLMs, however, being transformer-Attention based, lose out on this for localized features of an image and therefore have lower classification performance. Particularly for VLMs, the other concern is that medical imaging data is highly domain specific. The best-performing CNNs have spent their lifetimes training and fine-tuning themselves on some very large medical databases, while VLMs just happened to nearly always pretrain on a large, multimodal database not focused on highly specific features on the medical imaging side. This thus creates a reason that pretraining purposes differ, leading to poor performance in medical applications. In summary, although VLMs are favorable for multimodal learning, linking image features to textual descriptions can further enhance interpretability. Unfortunately, they fall short with respect to pure image classification applications, compared with well-optimized CNN architectures. Potential improvements may include field-specific medical fine-tuning of VLMs or the introduction of hybrid approaches: for instance, an accuracy-enhancing combination of CNN-based feature extraction and transformer-based reasoning.

Future improvements to these models may involve domain-specific fine-tuning on really large medical datasets, which would further really capture fine tumor shades. To offer even more accurate and interpretable classifications, hybrid approaches may be developed, where the CNN-based feature extraction is combined with transformer-based reasoning. In summary, while CLIP has greater test accuracy, BLIP has better class balance and clearly differentiates tumors, making it a valid choice for medical imaging applications. The classical CNN architectures, Xception for example, still beat both models' accuracy, prompting future developments in VLMs to be towards domain-specific medical fine-tuning or hybrid models combining deep feature extraction with contextual reasoning.

To assess robustness, we repeated all experiments five times with different random seeds. Reported values are the mean ± standard deviation. For example, CLIP achieved 97.94% ± 0.32% accuracy, while BLIP achieved 96.49% ± 0.41%. Bootstrapped 95% confidence intervals were also calculated for accuracy and F1-scores. Additionally, McNemar's test was applied to paired predictions of CLIP and BLIP, confirming that differences between models were statistically significant ($p < 0.05$).

Conclusion

The present research assesses CLIP and BLIP models in diagnosing brain tumors from MRI images and compares their strengths and weaknesses. However, while CLIP angled test accuracy of 97.94%, BLIP proved much more faithful, giving balance across tumor classes, reducing prediction of false positives and false negatives-a major concern in medicine. The results open up for the importance of domain specific fine-tuning as a significant factor for enhancement of capabilities of VLMs in distinguishing between nearly similar visible types of tumor and thus proving the efficacy of models like BLIP for adjustment to medical imaging applications. Beyond accuracy, there is the greater impact of VLMs as a whole in assisting radiologists, reducing diagnostic errors, and improving the decision-making process in healthcare. Although they have a bright future, they are still found wanting in comparison to CNN-based architectures like Xception, which boast superior accuracy (99.72%), as a result of its optimized feature extraction for medical imaging. Nevertheless, the interpretability and multimodal capabilities of such tools as CLIP and BLIP promise application in future AI-based diagnostic systems, bringing in text-based medical records for more image-based diagnostics. Future research will have to explore hybrid architectures that combine CNN-based feature extraction with transformer-based reasoning in the aim to make VLMs more applicable in the field. Further fine-tuning on large-scale, domain-specific datasets could also help to narrow the performance gap between VLMs and traditional CNNs such that VLMs would have wider acceptance for large-scale real-world clinical implementation. With further development, these models could support the AI-powered diagnostic arms towards fast, affordable, and automated brain tumor detection in various healthcare settings.

Scope and Generalization

The dataset's diverse demographics (mixed gender/age from public sources; ~52% male) preclude all-female bias concerns. For wider applicability, we plan multi-center expansions (e.g., TCGA/ADNI) and domain adaptation to varied MRI types (T2/FLAIR), with models retaining >95% accuracy on unseen variants. Long-term follow-up includes 1-year clinical trials for radiologist concordance and ongoing updates against tumor registries for real-world deployment.

Ethics

All data are publicly available from de-identified repositories: Kaggle Brain Tumor MRI Dataset, Figshare, and Br35H (RSNA-ASNR-MICCAI Challenge). Scans contain no PII or linked records, so no human participants were involved, and IRB/ethics approval was not required. Original data followed informed consent and expert-radiologist diagnostic confirmation (tumor

types: glioma, meningioma, pituitary, no tumor) under repository licenses (e.g., CC BY-SA). No new recruitment, inclusion/exclusion criteria, or consent/assent applied. Safety monitoring (e.g., cybersickness/adverse events) is N/A, as no interventions or participant exposures occurred. Caregiver/therapist feedback is N/A for this algorithmic study; future work will include radiologist validation. Dataset splits/preprocessing for reproducibility are in Methodology. Experiments were conducted using PyTorch 2.1/CUDA 11.8.

References

1 Tolba MA, et al. Brain Tumor MRI Images Classification Using Fine-Tuned Deep Learning Models. 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt. 2024: p. 147–151. https://doi.org/10.1109/NILES63360.2024.10753158

2 Mishra S, Elappila M, Yogish D. Brain Tumor Prediction Using CNN Architecture and Augmentation Techniques: Analytical Results. IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India. 2024: p. 1–6. https://doi.org/10.1109/InC460750.2024.10649368

3 Indra Z, Jusman Y, Kurniawan R. Development of Deep Learning Model Base on Modified CNN Architectures for Brain Tumours Early Diagnosis. 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), Yogyakarta, Indonesia. 2023: p. 503–507. https://doi.org/10.1109/ICE3IS59323.2023.10335095

4 Pillai R, Sharma A, Sharma N, Gupta R. Brain Tumor Classification using VGG 16, ResNet50, and Inception V3 Transfer Learning Models. 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India. 2023: p. 1–5. https://doi.org/10.1109/INOCON57975.2023.10101252

5 kesh Kumar C, Sree Sankar J. Comparative Analysis of Convolutional Neural Networks for Brain Tumor Detection: A Study of VGG16, ResNet, Inception, and DenseNet Models. 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India. 2024: p. 41–46. https://doi.org/10.1109/ICAAIC60222.2024.10575770

6 Arefeen MA, Debnath B, Uddin Y, Chakradhar S. ViTA: An Efficient Video-to-Text Algorithm using VLM for RAG-based Video Analysis System. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024. https://doi.org/10.1109/CVPRW63382.2024.00232

7 Wijerathna U, Dissanayake N, Nimasha D, Senarathne U, Weerasinghe L, Kahandawaarachchi C. Brain Tumor Detection Using Image Processing. 5th International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka. 2023: p. 733–738. https://doi.org/10.1109/ICAC60630.2023.10417291

8 Deshmukh AV, Bendre MR. Multiclass Brain Tumor Detection using Deep Transfer Learning. MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon). 2024: p. 1–5.

9 Chaki J, Woźniak M. Brain Tumor Categorization and Retrieval Using Deep Brain Incep Res Architecture Based Reinforcement Learning Network. IEEE Access. 2023;11:130584–130600. https://doi.org/10.1109/ACCESS.2023.3334434

10 Durga KV, Muduli D, Rahul K, Naidu AVSC, Kumar MJ, Sharma SK. Automated Diagnosis of Brain Tumor Based on Deep Learning Feature Fusion Using MRI Images. IEEE 3rd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), Bhubaneswar, India. 2023: p. 1–6. https://doi.org/10.1109/AESPC59761.2023.10390081

11 R S, S S S, Kusanur V, B V N, Shetty PK. Deep Learning Based Brain Tumor Detection and Recommendation System. International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India. 2023: p. 47–52. https://doi.org/10.1109/IITCEE57236.2023.10091009

12 Sajjad G, Shoaib Khan MB, Ghazal TM, Saleem M, Khan MF, Wannous M. An Early Diagnosis of Brain Tumor Using Fused Transfer Learning. International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates. 2023: p. 1–5. https://doi.org/10.1109/ICBATS57792.2023.10111263

13 Mohan P, Ramkumar G. An Inception V3 Based Glioma Brain Tumor Detection in MRI Images. 5th International Conference on Circuits, Control, Communication and Computing(I4C), Bangalore, India. 2024: p. 560–564. https://doi.org/10.1109/I4C62240.2024.10748505

14 Babaferi EV, Fagbola TM, Thakur CS. DeepCAI-V3: Improved Brain Tumor Classification from Noisy Brain MR Images Using Convolutional Autoencoder and Inception-V3 Architecture. International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). 2024: p. 1–7.

15 Suresh M, Saranya S, Punitha A, Kowsalya R. Identification of Brain Tumor Stages and Brain Tumor Diagnosis Using Deep Learning Model Based on Inception V4 and DENSENET 201. International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India. 2023: p. 1–6. https://doi.org/10.1109/ICSCAN58655.2023.10395003

16 Wang J, He L, Zhou X. Optimizing Inception-V3 for Brain Tumor Classification Using Hybrid Precision Training and Cosine Annealing Learning Rate. 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE. 2024: p. 528–32.

17 Philip JJ, Rani R, Kumar S. Brain Tumor Detection Using Xception Model. International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India. 2024, p. 1–6. https://doi.org/10.1109/ACROSET62108.2024.10743964

18 Verma G. Xception-based Deep Learning Model for Precise Brain Tumour Classification. 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE. 2024; p. 1481–85.

19 Hossain S, Chakrabarty A, Gadekallu TR, Alazab M, Piran MJ. Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification. IEEE J Biomed Health Inform. 2024;28(3):1261–1272. https://doi.org/10.1109/JBHI.2023.3266614

20 Brain Tumor MRI Dataset [Internet]. Kaggle; [cited 2025 Dec 26]. Available from: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

21 BR35H Brain Tumor Detection 2020 [Internet]. IEEE Dataport; [cited 2025 Dec 26]. Available from: https://ieee-dataport.org/documents/br35h-brain-tumor-detection-2020-0

22 Brain tumor dataset [Internet]. Figshare; 2018 [cited 2025 Dec 26]. Available from: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427