# Brain Matters: Enhancing Tumor Classification via CNN and Vision-Language Fusion

Chaudhari Khushi Ganesh
khushi21@iiserb.ac.in
Akshay Agarwal
akagarwal@iiserb.ac.in
Trustworthy BiometraVision Lab,
IISER Bhopal, India

**Abstract**

Accurate and efficient classification of brain tumors from Magnetic Resonance Imaging (MRI) is essential for timely diagnosis and treatment planning. Traditional methods often rely on hand-crafted features or vision-only deep learning models, which may not fully utilize the rich contextual information found in biomedical images. This work proposes a systematic hybrid framework for brain tumor classification that combines large Vision-Language Models (VLMs) with a ResNet backbone. Our goal is to enhance classification accuracy and reduce the diagnostic workload for radiologists. We tackle the challenge of multimodal information fusion by introducing a weighted concatenation mechanism to effectively merge features extracted from the VLM and the ResNet architecture. Additionally, we conducted initial zero-shot learning evaluations using prominent biomedical VLMs, BiomedCLIP and UniMedCLIP, to assess their inherent capabilities in medical image understanding without fine-tuning. This preliminary analysis helped inform the integration strategy for our proposed classification framework. Evaluated across six publicly available brain MRI datasets, our framework helps improve the limitations of zero-shot classification in medical vision-language models by synergistically combining CNN spatial features with semantic embeddings. This empirical benchmarking study enhances classification performance and has significant potential for streamlining the diagnostic workflow, ultimately easing the burden on medical professionals. One of the key goals of this research is to improve the zero-shot classification capability of medical vision-language models by leveraging spatial features from CNNs, thereby overcoming limitations seen in standalone zero-shot VLMs.

## 1 Introduction

Brain tumors are among the most critical and life-threatening neurological disorders, capable of impairing essential brain functions [27] depending on their type and anatomical location. Early and accurate diagnosis is vital, as it directly influences treatment planning, prognosis, and patient outcomes. Magnetic Resonance Imaging (MRI) remains the gold standard for brain tumor detection due to its superior spatial resolution and the ability to capture multicontrast tissue characteristics without the risks of ionizing radiation. However, MRI

interpretation depends highly on radiological expertise and is subject to variability between readers, especially in complex or borderline cases.

Recent advances in artificial intelligence (AI) have introduced scalable and objective tools for medical image analysis. Convolutional Neural Networks (CNNs), such as ResNet [11], have shown strong performance in extracting discriminative spatial features directly from raw MRI data [15]. More recently, Vision Transformers (ViTs) [9] have been explored for brain tumor classification, leveraging self-attention mechanisms to capture global contextual information in MRI scans [24]. However, their ability to capture a higher-level semantic context remains limited. On the other hand, vision-language models (VLMs), particularly domain-specialized variants such as BioMedCLIP [30] and UniMedCLIP [16], have emerged as powerful multimodal learners trained on large-scale biomedical image-text pairs. These models align visual content with clinical semantics, enabling zero-shot classification and more generalizable image interpretation without extensive retraining [29]. Despite this promise, their effectiveness in domain-specific tasks such as brain tumor classification remains limited, primarily due to a lack of fine-tuning on brain MRI datasets and insufficient spatial granularity.

To address these limitations, we propose a hybrid framework [1, 2] that effectively combines the strengths of CNNs and VLMs for robust brain tumor classification. We begin with a zero-shot evaluation of state-of-the-art VLMs, including LLaVA [18], UniMedCLIP [16], BioMedCLIP [30], and ChatGPT [5], which reveals their shortcomings in specialized neuroimaging tasks due to limited domain-specific fine-tuning. Building on these insights, we introduce a novel weighted feature fusion strategy that integrates spatial embeddings from ResNet-50 with semantically rich embeddings from BioMedCLIP. The fusion weights are derived from model performance metrics, ensuring interpretable and data-driven integration of complementary features. The resulting fused representations are then classified using machine learning models such as support vector machine [12], [20], decision tree classifier [22], and passive-aggressive classifier [8], yielding substantial performance gains over zero-shot of VLMs. This demonstrates the effectiveness of combining spatially grounded CNN features with semantically enriched VLM embeddings, improving generalizability and diagnostic accuracy. Importantly, our approach addresses the inherent limitations of zero-shot VLMs in brain tumor classification and highlights the potential of hybrid architectures to advance scalable and clinically relevant decision support systems in neuro-oncology.

## 2 Related Work

The rapid evolution of foundational models profoundly reshapes medical imaging, particularly for tasks like brain tumor classification, by offering robust generalization and zero-shot capabilities to address data scarcity [4, 25]. Recent advancements include adapting generalist vision models such as the Segment Anything Model (SAM) [17] for medical contexts, leading to specialized derivatives like MedicoSAM for image segmentation [3], and developing domain-specific foundational models exemplified by CheXFound for chest X-ray analysis [28]. Furthermore, the rapidly evolving field of Vision-Language Models (VLMs) represents a significant advancement, learning joint visual and textual representations to facilitate tasks such as zero-shot classification by aligning image features with text-based prompts. Prominent examples in this domain include CheXzero [23], BiomedCLIP, and UniMedCLIP, which are specifically adapted for biomedical image-text pairs. Moreover, medical VLMs such as MedCLIP [26] have reported strong performance on multimodal tasks, but

their application to neuroimaging remains underexplored. Unlike these, our work focuses on fusing domain-specialized BioMedCLIP embeddings with CNN spatial features for brain tumor MRI classification. Concurrently, broader multimodal models like LLaVA and specialized Large Language Models (LLMs) such as BioGPT [19] and general-purpose models like ChatGPT [5] have demonstrated powerful capabilities in generating and interpreting biomedical text or visual-language outputs. While these foundational models offer robust semantic understanding and remarkable generalization across various medical tasks, their direct application to highly specialized tasks such as brain tumor classification from high-resolution MRI data presents unique challenges. These include difficulties in capturing detailed spatial information and adapting models trained on general data to complex anatomical structures. Addressing these complexities, particularly in ensuring robust visual grounding and spatial accuracy for direct image analysis, remains a crucial area of ongoing investigation.

# 3 Proposed Algorithm for Brain Tumor Detection

We performed zero-shot classification to evaluate vision-language models (VLMs) capability in tumor detection. The models were used directly without fine-tuning or modifications, leveraging their multimodal pretraining on large-scale data. Some were general-purpose (e.g., LLaVA, ChatGPT), while others were domain-specific (e.g., BioMedCLIP, UniMedCLIP), trained on medical images and text. After observing the promising performance of BioMedCLIP and UniMedCLIP, we sought to make the approach more adaptable to our task by designing a hybrid pipeline focused on the vision-only component. This strategy avoids computationally expensive fine-tuning, typically requiring large-scale labelled data and significant resources.

We develop a hybrid deep learning framework that leverages spatial and semantic feature representations to address the limitations of zero-shot classification using vision-language models (VLMs) for brain tumor detection. The methodology includes dataset curation, feature extraction and fusion, and final classification. Several CNN backbones are explored, and ResNet-50 and ViT-L/16 are selected as spatial extractors, while BioMedCLIP and UniMedCLIP serve as vision-language models for semantic understanding. BioMedCLIP is pretrained on approximately 15M biomedical image–text pairs, capturing radiology-style descriptions and medical terminology. This domain-specific alignment is expected to aid MRI interpretation compared to general-purpose CLIP variants that lack medical semantics. All feature extractions are performed zero-shot to ensure generalization, without fine-tuning on the downstream task. The architecture incorporates three fusion strategies:

- **Concatenation of embeddings** – features from CNNs and VLMs are directly concatenated into a joint representation.
- **Score averaging from independent classifiers** – prediction scores from separate classifiers are averaged to obtain the final decision.
- **Weighted fusion using validation precision scores** – classifier outputs are combined with learned weights based on validation performance.

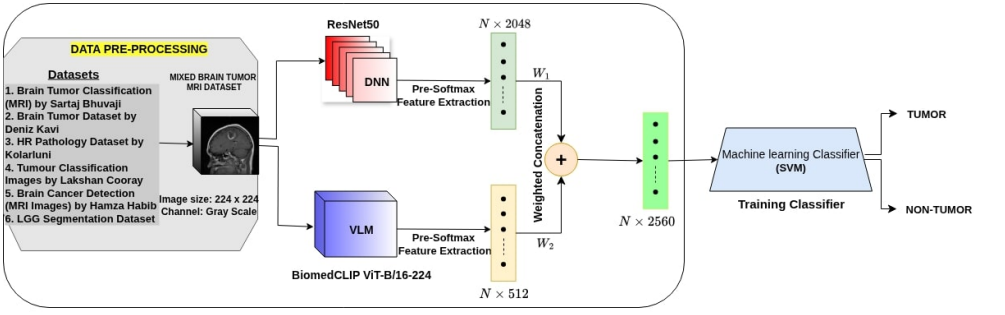



Figure 1: Proposed VLM-CNN hybrid model for brain tumor classification. The pipeline integrates spatial features from ResNet-50 and semantic features from BioMedCLIP, followed by weighted fusion and traditional classification.

Let $\mathbf{f}_{\text{resnet}} \in \mathbb{R}^{d_1}$ and $\mathbf{f}_{\text{clip}} \in \mathbb{R}^{d_2}$ denote the feature embeddings extracted from ResNet-50 and BioMedCLIP, respectively. We define a weight vector $\mathbf{w} = [w_1, w_2]$ and compute normalized fusion weights as:

$$\alpha_i = \frac{e^{w_i}}{e^{w_1} + e^{w_2}}, \quad i = 1, 2 \tag{1}$$

The fused feature representation is obtained by concatenating the weighted embeddings:

$$\mathbf{f}_{\text{fused}} = \left[ \alpha_1 \cdot \mathbf{f}_{\text{resnet}} \parallel \alpha_2 \cdot \mathbf{f}_{\text{clip}} \right] \in \mathbb{R}^{d_1 + d_2} \tag{2}$$

where $\parallel$ denotes concatenation along the feature dimension.

The fused representation $\mathbf{f}_{\text{fused}}$ is then used for classification with a traditional machine learning model, such as an SVM:

$$\hat{y} = \text{Classifier}(\mathbf{f}_{\text{fused}}), \quad \text{Classifier} \in \{\text{SVM, Logistic Regression, Passive Aggressive, etc.}\} \tag{3}$$

The fused feature $\mathbf{f}_{\text{fused}}$ is directly used as input to a traditional machine learning classifier. This decoupling from end-to-end neural training enhances interpretability and mitigates overfitting, though it may limit the model's capacity to learn more complex joint representations. The empirically determined optimal weights are $\mathbf{w}_{\text{resnet}} = 0.47$ and $\mathbf{w}_{\text{biomed}} = 0.53$, based on validation accuracy. After fusion, the final classification is performed using interpretable traditional classifiers such as SVM, Logistic Regression, Passive Aggressive, and SGD, with the best-performing model selected. Although weighted fusion is straightforward, we adopt it for its interpretability and reduced risk of overfitting.

The proposed Hybrid CNN-VLM Classifier integrates spatial and semantic representations for robust brain tumor classification. As illustrated in Figure 1 and the pseudocode of Algorithm 1, MRI images are first preprocessed through resizing and normalization, after which spatial features are extracted using ResNet-50 and semantic embeddings are obtained from BioMedCLIP. These complementary features are fused via a weighted concatenation mechanism, where modality-specific weights are normalized to ensure balanced contributions. The fused representation is then directly used for classification with a traditional machine learning model such as a support vector machine (SVM). This modular architecture

**Algorithm 1** Proposed Hybrid (VLM + CNN) Classifier for Brain Tumor

**Require:** Input: MRI image $x$
**Ensure:** Output: Label $y \in \{\text{Tumor}, \text{Non-Tumor}\}$
1. Preprocess $x$ (resize, normalize)
2. $f_{\text{ResNet}} \leftarrow \text{ResNet50}(x)$
3. $f_{\text{BioMed}} \leftarrow \text{BioMedCLIP}(x)$
4. Compute fusion weights: $\alpha_1, \alpha_2$
5. $f_{\text{fused}} \leftarrow [\alpha_1 \cdot f_{\text{ResNet}} \parallel \alpha_2 \cdot f_{\text{BioMed}}]$
6. $y \leftarrow \text{Classifier}(f_{\text{fused}})$
7. **return** $y$

leverages the strengths of CNN-driven spatial detail and VLM-derived semantic context to improve diagnostic accuracy, while maintaining flexibility to integrate additional modalities (e.g., clinical metadata) and supporting transparent decision-making, an essential requirement in medical imaging applications.

## 3.1 Tumor Classification Results and Analysis

We construct a large-scale dataset by aggregating six publicly available brain MRI repositories:

- The Cancer Genome Atlas Low Grade Glioma Collection [21]
- Brain Tumor Classification (MRI) [6]
- Brain Tumor Image Dataset [7]
- MRI Braintumor Glioma Dataset[14]
- Tumour Classification Images [13]
- Brain Cancer Detection MRI Images [10]

All six datasets are provided in standard image formats (JPEG/PNG) with consistent dimensions and orientation, pre-aligned for classification tasks by the original providers. While the datasets have been somewhat standardized, potential label heterogeneity and batch effects may remain due to their diverse sources. Minimal preprocessing, such as resizing and format conversion, was applied. Minimal preprocessing, such as resizing and PNG conversion, enables effective embedding extraction for zero-shot and fusion-based models. We uniformly resize all MRI images to ensure compatibility across sources and convert them into PNG format. No further preprocessing, such as intensity normalization, skull-stripping, or spatial re-alignment, is necessary. Label harmonization maps all tumor subtypes (gliomas, meningiomas, pituitary tumors, etc.) into a single Tumor class. At the same time, non-tumor scans are assigned to the Non-Tumor class, yielding a binary classification setup. The final dataset is balanced, with the training set containing 5,931 tumor and 5,587 non-tumor images, the test set containing 5,143 tumor and 5,056 non-tumor images from which only 4000 are used to test the model, and a separate validation set used for model selection. This split allows us to tune model fusion weights on the validation set while evaluating zero-shot and hybrid CNN+VLM performance on a fully held-out test set, ensuring unbiased assessment.

This streamlined approach simplifies the workflow without compromising performance, as binary classification (tumor vs. non-tumor) is relatively robust to residual appearance variation, allowing our hybrid CNN+VLM framework to generalize effectively with minimal preprocessing overhead.

We conduct multiple experiments to arrive at our conclusions, comprising two main components: evaluating the zero-shot performance of Vision-Language Models (VLMs) and assessing the supervised performance of our proposed hybrid CNN-VLM fusion pipeline. All experiments are conducted on a workstation with an Intel Core i7-14700 CPU and NVIDIA GeForce RTX 4070 Ti GPU. Our fusion framework is implemented in PyTorch 2.6.0 with CUDA 11.8. All experiments use the Adam optimizer with a learning rate of $1 \times 10^{-3}$, batch size of 64, and 50 training epochs. A validation split comprising 20% of the available data is used to estimate fusion weights, which are then fixed for final training and evaluation. The fused representations are subsequently used to train an SVM classifier, and performance is reported on the held-out test set.

## 3.2 Zero-shot Analysis

We evaluate four vision-language models to classify brain tumors in a zero-shot setting: BioMedCLIP, UniMedCLIP, LLaVA, and ChatGPT. In zero-shot inference, no model parameters are fine-tuned; instead, we directly exploit their pretrained vision–language alignment. For CLIP-based models (BioMedCLIP, UniMedCLIP), tumor classification is performed by computing cosine similarity between MRI image embeddings and text prompts describing tumor classes (e.g., "MRI of a brain with tumor" vs. "MRI of a normal brain"). For generative models (ChatGPT, LLaVA), we provide descriptive diagnostic prompts and interpret their text-based predictions. Accuracy for each model is computed as the ratio of correctly predicted cases (true positives + true negatives) to the total number of test samples. As shown in Table 1, BioMedCLIP delivers the most balanced performance, particularly on high-grade tumors. UniMedCLIP is highly sensitive to tumors but often misclassifies normal cases, whereas ChatGPT and LLaVA perform inconsistently due to their lack of domain-specific visual grounding. These results highlight the limitations of zero-shot models and the need for a supervised hybrid approach.

Table 1: Zero-shot performance of VLMs on the test MRI dataset.

| Model | Overall Acc. | Non-tumor (0) | Tumor (1) |
|---|---|---|---|
| BioMedCLIP | **62.05%** | 66.71% | 57.66% |
| UniMedCLIP | 48.92% | 1.07% | **91.81%** |
| ChatGPT | 51.00% | 32.0% | 70.0% |
| LLaVA | 50.43% | 28.0% | 72.8% |

BioMedCLIP, while the most reliable among evaluated models, achieves moderate accuracy ( 62.05%) and exhibits room for improvement. The zero-shot VLM performances for specialized neuroimaging tasks remain limited, motivating our hybrid approach. UniMedCLIP, while highly tumor-sensitive, struggles with non-tumor class detection, suggesting bias. ChatGPT and LLaVA, lacking domain-specific visual grounding and confidence outputs, produce inconsistent results with higher false positives, especially in subtle cases.

Table 2: Model-specific caption prompts for output generation.

| Model | Caption |
|---|---|
| BioMedCLIP(Caption classes) | This brain MRI shows a tumor. |
| | This brain MRI is normal without a tumor. |
| UniMedCLIP(Caption classes) | Brain MRI scan with tumor. |
| | Brain MRI scan without tumor. |
| LLaVA | Are these healthy scans? Give an answer in 1 word |
| | Can you detect any abnormality in this MRI scan? |
| ChatGPT | Can you detect any abnormality in the provided image? |
| | Is there a tumor present in this brain MRI scan? |

Table 2 presents examples of caption classes provided to the models. These are not an exhaustive set of captions used, but illustrative samples to demonstrate that the model's output varies depending on the chosen caption formulation.

## 3.3 Results of the Hybrid Model

This section evaluates our hybrid classification framework that integrates convolutional and vision-language models. We test various model pairings and fusion strategies, demonstrating that the final combination of ResNet-50 and BioMedCLIP with weighted fusion achieves the best performance.

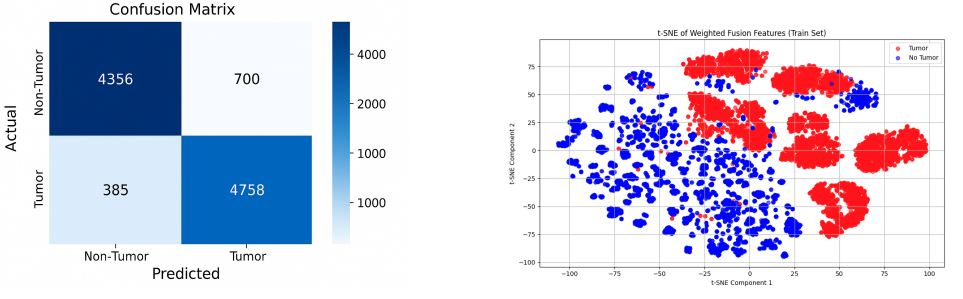
To address the limitations of zero-shot classification, we design a supervised pipeline that combines spatial features from ResNet-50 with semantic features from BioMedCLIP. Features from both models are fused using three strategies: (i) feature concatenation, (ii) score averaging, and (iii) weighted fusion based on validation precision. We conduct comprehensive experiments with different model pairs, including ViT, ResNet, UniMedCLIP, and BioMedCLIP. As shown in Table 3, the ResNet-50 + BioMedCLIP combination consistently outperforms others across all fusion strategies. Notably, the weighted fusion approach with this pair achieves the highest validation accuracy of **89.3%**, demonstrating its effectiveness in combining spatial and semantic modalities. We assess whether our fusion framework leads to tangible improvements over zero-shot VLM baselines, thus addressing critical gaps in medical image understanding.

The performance of the proposed ResNet + BioMedCLIP weighted fusion approach is evaluated using the SVM classifier. The model achieves an overall accuracy of 89.3% and an F1 Score of 89.3%. Sensitivity, which measures the proportion of tumor scans correctly identified, is 88.9%, while specificity, reflecting the proportion of non-tumor scans correctly classified, is 89.7%. The ROC-AUC, representing the model's ability to discriminate between tumor and non-tumor classes across varying thresholds, is 89.3%. These results demonstrate that the fused feature representation provides robust and balanced performance across both classes, highlighting the effectiveness of the hybrid model.

To further validate the model's effectiveness, we visualise the fused feature space using a confusion matrix and a t-SNE plot, as shown in Figure 2. The confusion matrix illustrates balanced classification between tumor and non-tumor categories, while the t-SNE projection confirms clear class separation, highlighting the discriminative strength of the learned embeddings. Together with the quantitative results, these visualizations support the robustness and interpretability of our hybrid ResNet-50 + BioMedCLIP model.

Table 3: Validation accuracy (%) and F1-score (%) for different model combinations and fusion strategies. The best performance is achieved using ResNet + BioMedCLIP with weighted fusion.

| Model Pair | Fusion Strategy | Accuracy (%) | F1-score (%) |
|---|---|---|---|
| ResNet + UniMedCLIP | Concatenation | 77.0 | 76.0 |
| ResNet + UniMedCLIP | Score Averaging | 67.1 | 67.0 |
| ResNet + UniMedCLIP | Weighted Fusion | 88.3 | 88.3 |
| ViT + UniMedCLIP | Concatenation | 75.1 | 74.0 |
| ViT + UniMedCLIP | Score Averaging | 70.9 | 71.0 |
| ViT + UniMedCLIP | Weighted Fusion | 88.0 | 88.3 |
| ViT + BioMedCLIP | Concatenation | 82.0 | 82.0 |
| ViT + BioMedCLIP | Score Averaging | 83.4 | 83.0 |
| ViT + BioMedCLIP | Weighted Fusion | 82.2 | 82.1 |
| ResNet + BioMedCLIP | Concatenation | 87.3 | 87.0 |
| ResNet + BioMedCLIP | Score Averaging | 83.3 | 83.0 |
| **ResNet + BioMedCLIP** | **Weighted Fusion** | **89.3** | **89.3** |




(a) Confusion matrix

(b) t-SNE plot of fused features

Figure 2: Visualisation of the proposed hybrid model performance.

## 3.4 Discussion

Our experiments confirm that the proposed hybrid pipeline substantially boosts zero-shot classification capability, demonstrating its value for medical tasks where labelled data is limited. Our findings highlight the potential and limitations of vision-language models (VLMs) in medical imaging. Although BioMedCLIP and UniMedCLIP demonstrate promise in zero-shot tumour classification, they struggle with class imbalance and lack fine-grained spatial understanding, most notably UniMedCLIP, which heavily favours tumour cases at the cost of normal detection. General-purpose models like ChatGPT and LLaVA underperform due to weak domain-specific visual grounding and over-reliance on language priors. These challenges motivate our hybrid approach, where combining ResNet-50's spatial features with BioMedCLIP's semantic embeddings provides a more balanced representation. The superior performance of weighted fusion supports the view that modular integration of complementary models offers a more scalable and interpretable alternative to monolithic, end-to-end systems in medical AI. However, improving generalizability, clinical interpretability, and robustness to imaging variability remains an open challenge. While binary tumour vs.

non-tumour classification provides proof-of-concept, it is less clinically useful than tumour grading, subtype differentiation, or segmentation required in practice.

### 3.5 Limitations

This study demonstrates strong binary classification performance, but merging diverse tumour subtypes into a single class limits clinical applicability. While our softmax-normalised fusion weights provide interpretable modality contributions, they remain fixed and may restrict the capacity to learn richer joint representations.

## 4 Conclusion

We presented a hybrid deep learning framework for binary brain tumor classification using MRI, combining the spatial capabilities of CNNs with the semantic power of vision language models. Zero-shot analysis revealed the limitations of standalone VLMs, particularly in medical imaging contexts lacking domain-specific fine-tuning. Our proposed architecture, which fuses ResNet-50 and BioMedCLIP features via a weighted strategy, outperforms individual models and simpler fusion methods. These results demonstrate the effectiveness of combining complementary representations for robust medical image analysis. In the future, we aim to advance zero-shot brain tumor image analysis and ensure no deadly cases are left behind for diagnosis.

## References

[1] Akshay Agarwal and Nalini Ratha. Deepfake catcher: Can a simple fusion be effective and outperform complex DNNs? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 3791–3801, 2024.

[2] Akshay Agarwal, Afzel Noore, Mayank Vatsa, and Richa Singh. Generalized contact lens iris presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):373–385, 2022.

[3] Anwai Archit, Luca Freckmann, and Constantin Pape. Medicosam: Towards foundation models for medical image segmentation. *arXiv preprint arXiv:2501.11734*, 2025.

[4] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[5] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. Chatgpt: Applications, opportunities, and threats. In *IEEE Systems and Information Engineering Design Symposium*, pages 274–279, 2023.

[6] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri), 2020. URL https://www.kaggle.com/dsv/1183165.

[7] Jun Cheng. Brain tumor dataset. https://doi.org/10.6084/m9.figshare.1512427.v5, 2017.

[8] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar): 551–585, 2006.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Hamza Habib. Brain cancer detection mri images. https://www.kaggle.com/datasets/hamzahabib47/brain-cancer-detection-mri-images, 2025.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[13] Somasundaram Karuppanagounder and Kalavathi Palanisamy. Medical image contrast enhancement based on gamma correction. *International Journal of Knowledge Management and e-Learning*, 3:15–18, 05 2011.

[14] Kolar Khan. Mri_braintumor_glioma_dataset. https://www.kaggle.com/dsv/6381814, 2023.

[15] Md Saikat Islam Khan, Anichur Rahman, Tanoy Debnath, Md Razaul Karim, Mostofa Kamal Nasir, Shahab S Band, Amir Mosavi, and Iman Dehzangi. Accurate brain tumor detection using deep convolutional neural network. *Computational and structural biotechnology journal*, 20:4733–4745, 2022.

[16] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[19] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[20] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.

[21] N. Pedano, A. E. Flanders, L. Scarpace, T. Mikkelsen, J. M. Eschbacher, B. Hermes, V. Sisneros, J. Barnholtz-Sloan, and Q. Ostrom. The cancer genome atlas low grade glioma collection (tcga-lgg) (version 3). https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK, 2016. [Data set].

[22] Philip H. Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.

[23] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

[24] Sudhakar Tummala, Seifedine Kadry, Syed Ahmad Chan Bukhari, and Hafiz Tayyab Rauf. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29(10):7498–7511, 2022.

[25] Vivien van Veldhuizen, Vanessa Botha, Chunyao Lu, Melis Erdal Cesur, Kevin Groot Lipman, Edwin D de Jong, Hugo Horlings, Clárisa Sanchez, Cees Snoek, Ritse Mann, et al. Foundation models in medical imaging–a review and outlook. *arXiv preprint arXiv:2506.09095*, 2025.

[26] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*, page 3876, 2022.

[27] Archana Yadav, Vishakha Pareek, Akshay Agarwal, and Santanu Chaudhury. Neural encoding of odors: Translating odors into unique digital representation with eeg signals. In *International Conference on Pattern Recognition*, pages 280–295. Springer, 2024.

[28] Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Chest x-ray foundation model with global and local representations integration. *arXiv preprint arXiv:2502.05142*, 2025.

[29] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[30] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.