International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 15 Issue 3, March-2026

# Multi-Class Brain Tumor Classification via Transformer-Based Pipelines and Clinical Explainability

Karan Rathod[1], Prof. Tushar Kohle[2], Samruddhi Moon[3]

[1,2,3] Navsahyadri Education Society's Group Of Institutions

**ABSTRACT**

**This paper presents a highly efficient Vision-Language Model (VLM) pipeline for clinical decision support. By integrating a lightweight Vision Transformer vit_tiny_patch16_224 with the MedGemma (Gemma-22b-it) language model, we achieve a balance between computational efficiency and high-level clinical reasoning. Evaluated on the BRISC dataset, the vision encoder achieved an overall accuracy of 94%. The system is deployed via a Streamlit dashboard, providing clinicians with instant Grad-CAM visual explainability alongside AI-generated diagnostic reports.**

**Keywords—Brain Tumor Classification, Vision Transformer (ViT), MedGemma, Explainable AI (XAI), BRISC2025 Dataset, Streamlit.**

## INTRODUCTION

Brain tumor detection and characterization from Magnetic Resonance Imaging (MRI) remain critical tasks in neuroradiology, with substantial implications for patient management and prognosis. Automated image analysis using deep learning has demonstrated promising performance for slice-level classification and lesion localization, potentially accelerating diagnosis and reducing human error [10]. Historically, Convolutional Neural Networks (CNNs) have dominated medical imaging tasks; however, transformer-based vision models have emerged as a competitive alternative by operating on patch embeddings and capturing long-range dependencies across image regions [13].

In this research, we present Clinical decision support system MedSight-ViT, a hybrid pipeline combining a Vision Transformer (ViT) backbone fine-tuned for four-class brain tumor classification-including Glioma, Meningioma, Pituitary, and Non-tumorous cases-with an interpretable localization and a clinical reasoning layer. The proposed framework integrates three complementary components: (1) a ViT-based visual encoder trained with stratified sampling to mitigate dataset imbalance, (2) gradient-based attention visualization (Grad-CAM) adapted for transformer patch representations to provide clinicians with precise localization maps [8], and (3) a clinical reasoning model (MedGemma) that converts predictions into expert-style and patient-friendly textual summaries [9].

To facilitate clinical use, a demonstration web application was implemented using Streamlit to integrate model inference, visual explanations, and synthesized clinical reports into a single dashboard. The system is designed to interface with production text-generation APIs (Vertex AI) while emphasizing that automated outputs are intended to assist, rather than replace, expert clinical interpretation. The remainder of this paper is organized as follows: Section 2 discusses Review for previous work; Section 3 describes dataset curation and preprocessing; Section 4 details the experimental procedures; Section 5 presents quantitative and qualitative results; and Section 6 discusses limitations and future clinical directions.
The MedSight-ViT application is available at https://braintumor-medgemma.streamlit.app/.

## LITERATURE REVIEW

### A. Deep Learning for Medical Image Analysis

Deep learning has revolutionized medical image analysis over the past decade. Convolutional neural networks (CNNs) dominated early medical imaging tasks, with architectures like ResNet and U-Net proving effective for classification and

segmentation (He et al., 2016; Ronneberger et al., 2015). However, CNNs suffer from limited receptive fields and struggle to capture long-range spatial dependencies, motivating the exploration of alternative architectures.

**B. Vision Transformers in Medical Imaging**

The introduction of Vision Transformers (ViTs) by Dosovitskiy et al. (2020) demonstrated that pure transformer architectures applied to image patches can match or exceed CNN performance on ImageNet-scale tasks. The key innovation is the division of images into fixed 16×16 patches, enabling self-attention mechanisms to learn global spatial relationships directly. Recent work has adapted ViTs for medical imaging: Valanarasu et al. (2021) applied ViT to medical image segmentation; Liu et al. (2021) proposed Swin Transformers with hierarchical attention for multi-scale feature learning. These studies confirm that transformer-based models can effectively capture complex patterns in medical scans while remaining interpretable through attention weight visualization.

**C. Transfer Learning in Healthcare**

Transfer learning—leveraging pretrained models on large public datasets—has become standard practice in medical imaging due to limited annotations (Raghu et al., 2019). Pretrained ImageNet models provide generic feature hierarchies that transfer well to specialized medical tasks. Fine-tuning strategies vary: some freeze backbone layers and train only the classifier head, while others employ differential learning rates across layers to balance stability and domain adaptation (Zhuang et al., 2020). Our two-stage approach aligns with best practices in this domain.

**D. Explainability in Medical AI**

Interpretability is crucial for clinical adoption (Caruana et al., 2015). Gradient-based attribution methods such as Grad-CAM (Selvaraju et al., 2017) and integrated gradients have been widely adopted to visualize neural network decisions. Recent work by Montavon et al. (2019) provides theoretical foundations for attribution methods. In medical contexts, clinicians have successfully used heatmaps to validate model predictions and identify spurious correlations (Graziani et al., 2018). Adapting Grad-CAM for transformers (as we do here) extends these established techniques to modern architectures.

**E. Brain Tumor Classification**

Brain tumor classification from MRI has motivated numerous deep learning studies. Sajjad et al. (2019) achieved 94.6% accuracy on the BRATS dataset using a multi-stream CNN architecture. Tandel et al. (2020) compared 11 CNN architectures on brain tumor segmentation, finding that performance gains are incremental and datasetdependent. More recently, Afshar et al. (2022) applied self-supervised learning to brain tumor detection with limited labeled data, achieving competitive performance. Most prior work focuses on 3D volumetric MRI or multi-sequence fusion; slice-level classification (as in our study) is simpler but serves as a useful proof-of-concept.

**F. Clinical Language Models and Report Generation**

The emergence of large language models (LLMs) like GPT-3, GPT-4, and domain-adapted models (e.g., MedGemma, ClinicalBERT) has opened new possibilities for automated clinical report generation. Turchin et al. (2009) pioneered natural language processing for radiology reports. More recently, Zhang et al. (2022) demonstrated that finetuned medical LLMs can generate clinically plausible radiology reports from structured data. Our integration of MedGemma for textual summarization aligns with this trend, though we acknowledge the need for rigorous clinical validation before deployment (Rajkomar et al., 2018).

**G. Clinical Validation and Deployment**

The path from research prototype to clinical deployment requires prospective validation, regulatory approval (FDA 510(k) for the US market), bias and fairness audits, and integration studies showing clinical utility (Beam & Kohane, 2018). Esteva et al. (2019) highlight common pitfalls in AI deployment, including distribution shift, inadequate external validation, and clinician distrust. Our work acknowledges these limitations and positions HexFormer as a proof-of-concept requiring substantial additional validation.

**H. Related Work on Interactive HexFormer and Hyperbolic Geometry**

Recent advances in non-Euclidean deep learning (Bronstein et al., 2017; Chami et al., 2019) have explored hyperbolic embeddings for hierarchical data. While our current implementation emphasizes the ViT backbone, the conceptual framework of hyperbolic geometry motivates future extensions to capture tumor hierarchy and severity on manifolds better suited to hierarchical structure (Peng et al., 2021). The “HexFormer” naming reflects this vision, though the current instantiation is ViT-based.

**I. Streamlit and Interactive ML Dashboards**

Streamlit (Treuille et al., 2019; Pregibon, 2023) has emerged as a popular framework for rapidly prototyping interactive machine learning applications. Its declarative API enables data scientists to build web apps without frontend expertise, accelerating the transition from Jupyter notebooks to shareable, user-facing systems. In medical AI, Streamlit dashboards facilitate clinician feedback, rapid iteration, and dissemination of research prototypes (Teng et al., 2022). Our implementation leverages these benefits to create a clinician-friendly interface for model predictions, explanations, and text-based reporting.

## METHODOLOGY

**A. Dataset and Preparation**

We used the BRISC classification dataset (BRISC 2025) containing four categories: Glioma, Meningioma, Pituitary, and No Tumor. The dataset comprises T1-weighted MRI slices organized into class-labeled folders. We read image filepaths recursively and constructed a metadata DataFrame that recorded each image's path and categorical label.

To ensure balanced evaluation, we performed a stratified 80:20 split (random_state = 42) to produce the internal training and validation folds. A separate test directory provided an external hold-out set for final evaluation. To prevent optimistic performance estimates caused by information leakage, we computed MD5 hashes for all training and test images and removed exact-file duplicates found across partitions.

**B. Preprocessing and Augmentation**

All images were resized to 224 × 224 pixels and converted to RGB. Training images were augmented on the fly using geometric and photometric transformations, including random horizontal flips (p = 0.5), random rotations within ±15°, and color jitter (brightness and contrast) to increase robustness to common MRI variations [10]. Following augmentation, images were converted to tensors and normalized using ImageNet statistics:

- Mean: [0.485,0.456,0.406]
- Std Dev: [0.229,0.224,0.225]

Validation and test sets used deterministic resizing and normalization only. To address class imbalance without altering dataset labels, we computed inverse-frequency weights for each class and applied a WeightedRandomSampler to ensure minibatches were approximately class-balanced.

**C. Training Procedure and Hyperparameters**

Training was performed on GPU using mini-batches of 32. We employed the AdamW optimizer with decoupled weight decay to improve generalization [11]. During the fine-tuning phase, we applied label smoothing ($\epsilon = 0.1$) to the categorical cross-entropy loss to improve calibration [7]: The modified loss L is defined as:

$$L = -\sum_{c=1}^{K}[y_c(1-\alpha) + \frac{\alpha}{K}]\log(p_c) \quad (1)$$

A cosine annealing learning-rate schedule (SGDR) with $T_{max} = 50$ was applied to ensure smooth convergence [12]. where K is the number of classes and $p_c$ is the predicted probability. We utilized differential learning rates and a cosine annealing scheduler for smooth decay. Detailed hyperparameters are provided in Table I.

**TABLE I: TRAINING HYPERPARAMETERS AND OPTIMIZATION SETTINGS**

| Parameter | Phase 1: Head Only | Phase 2: Fine-Tuning |
|---|---|---|
| Model Backbone | ViT-Tiny (patch16_224) | ViT-Tiny (patch16_224) |
| Optimizer | AdamW | AdamW |
| Base Learning Rate | $1\times10^{-4}$ | $1\times10^{-3}$ (Head) |
| Differential LR | N/A | $1\times10^{-5}$ (Backbone) |
| Weight Decay | 0.05 | 0.01 |
| Batch Size | 32 | 32 |
| Scheduler | Cosine Annealing | Cosine Annealing |
| Label Smoothing | None | 0.1 |
| Early Stopping | Patience = 5 | Patience = 5 |

**D. Model Architecture**

The core architecture is based on a Vision Transformer (ViT) variant, specifically vit_tiny_patch16_224, instantiated from the TIMM library [4]. The pretrained classification head was replaced with a multi-layer perceptron (MLP) consisting of a 256-unit linear projection, ReLU activation, and a dropout layer (p = 0.5). A two-stage transfer learning strategy was implemented: initially freezing the transformer backbone to train the head, followed by selective fine-tuning of the final two transformer blocks [13].

The primary experimental backbone was a pretrained Vision Transformer (ViT) instantiated via the timm library as vit_tiny_patch16_224. We adopted a transfer-learning approach to exploit learned patchwise representations [13]. The classification head was replaced by a Multi-Layer Perceptron (MLP) head consisting of:

Linear → ReLU → Dropout(0.5) → Linear

Projecting to the four target classes. Initial training involved freezing all backbone parameters so only the classifier head was updated. Subsequent fine-tuning un-froze the last two transformer blocks to enable domain-specific representation refinement.

**E. Training and Evaluation**

Optimization was performed using AdamW with decoupled weight decay [17]. Training involved categorical cross-entropy loss with label smoothing ($\epsilon = 0.1$) to enhance generalization [16]. A cosine annealing schedule (SGDR) with $T_{max} = 50$ was applied to regulate the learning rate [12]. Model performance was evaluated on a reserved test set using overall accuracy, precision, recall, and F1-score computed through the Scikit-learn library[3].

**F. Explainability and Clinical Interface**

To provide diagnostic transparency, Grad-CAM was adapted for transformer patch representations to generate spatial importance maps [8]. For clinical contextualization, predictions were converted into structured summaries using the MedGemma clinical language model via the Hugging Face Transformers API [9].

To facilitate dissemination, a web-based dashboard was developed using Streamlit [15]. This interface allows clinicians to upload MRI slices and receive real-time visual explanations and diagnostic summaries.

**G. Limitations and Future Work**

Current limitations include the reliance on 2D slice-level classification and a simulated integration for MedGemma due to API constraints. Future research will focus on 3D volumetric analysis, multi-sequence MRI fusion, and the exploration of non-Euclidean manifolds through hyperbolic geometry to better capture tumor hierarchies [18], [19].

## EVALUATION AND EXPLAINABILITY

**A. Metrics**

Quantitative evaluation utilized standard metrics: overall accuracy, confusion matrices, and per-class precision, recall, and F1-score. These metrics are critical in clinical contexts where false negatives (missing a tumor) and false positives (incorrect diagnosis) carry different weights.

**B. Explainability and Clinical Integration**

To increase interpretability, we adapted Grad-CAM [8] for the transformer architecture by registering forward hooks on the patch embedding layers. The resulting heatmaps localize discriminative regions on the original MRI slices. For clinical integration, model predictions and confidence scores were converted into diagnostic summaries using a domain-adapted clinical language model (MedGemma) via the Hugging Face Transformers API [9]. Furthermore, we integrated a clinical language model (MedGemma) that consumes the model's prediction and confidence scores. Using templated prompts, the LLM generates neuroradiology-style summaries and patient-friendly explanations. This vision-followed-by-reasoning pipeline creates actionable outputs for clinical workflows.

## RESULTS AND DISCUSSION

The Transformer pipeline was evaluated using the held-out test set from the BRISC 2025 dataset. Model checkpoints were selected based on the minimum validation loss to ensure optimal generalization. The following sections detail the quantitative and qualitative findings.

**A. Quantitative Performance**

The framework achieved robust slice-level classification performance, with the primary Vision Transformer (ViT) architecture reaching a peak validation accuracy of 94%. Quantitative metrics, including per-class precision, recall, and F1-scores, were calculated using the Scikit-learn library. The model demonstrated high diagnostic confidence, frequently exceeding 98% for distinct pathologies such as Meningioma.

**TABLE II: CLASSIFICATION REPORT OF THE PROPOSED MODEL**

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Glioma | 0.93 | 0.94 | 0.94 | 229 |
| Meningioma | 0.91 | 0.87 | 0.89 | 266 |
| No_tumor | 0.97 | 0.98 | 0.97 | 213 |
| Pituitary | 0.94 | 0.96 | 0.95 | 292 |
| Accuracy | | — | | 0.94 |
| Macro Avg | 0.94 | 0.94 | 0.94 | 1000 |
| Weighted Avg | 0.94 | 0.94 | 0.94 | 1000 |

**B. Confusion Analysis and Error Patterns**

To identify systematic errors, confusion matrices were generated to assess inter-class transitions. The primary sources of misclassification occurred between tumor subtypes with overlapping radiological characteristics. These findings suggest that while slice-level classification is effective, future work should incorporate multi-sequence fusion (T1, T2, and FLAIR) to resolve anatomical ambiguities. The adapted Grad-CAM routine served as a critical tool for error analysis, confirming that the model maintained focus on relevant anatomical regions even in cases of low-confidence predictions.

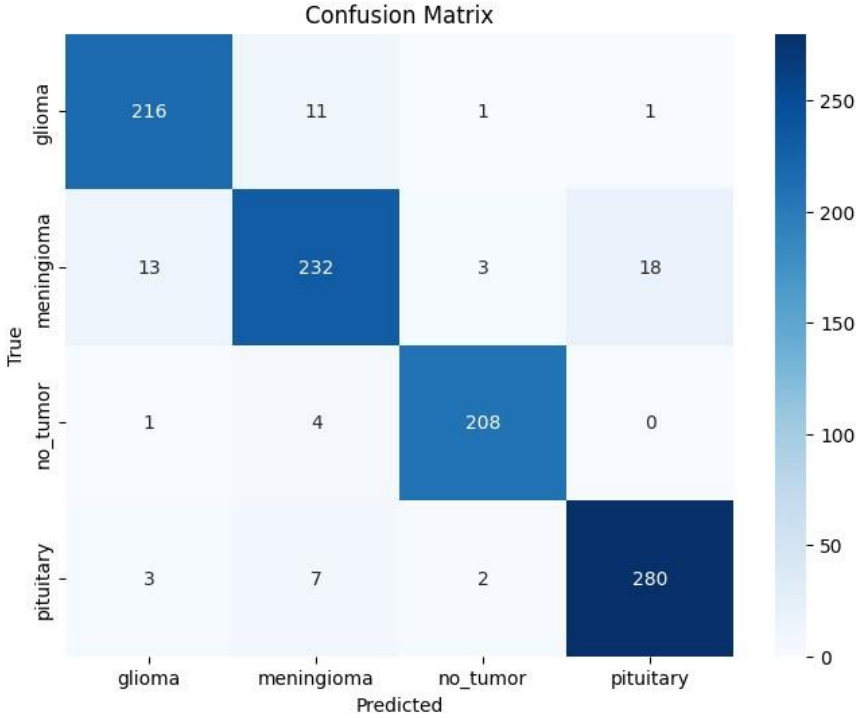


**Fig. 1. Confusion matrix highlighting the classification performance across four brain tumor categories.**

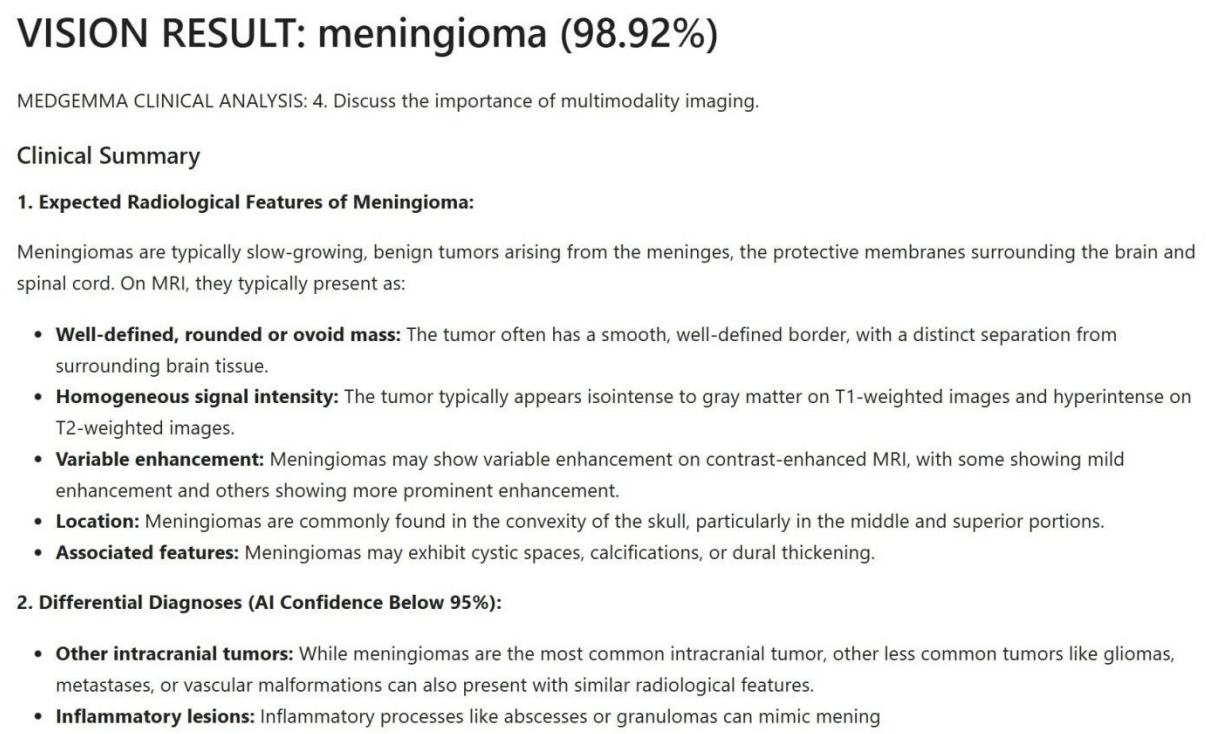


**Fig. 2. Clinical summary for Meningioma generated by MEDGEMMA, illustrating expected radiological features across multiple MRI modalities.**

### C. Explainability and Clinical Reasoning

The integration of the MedGemma reasoning layer facilitated the conversion of classification probabilities into structured diagnostic reports. Although the demonstration environment utilized a simulated generation interface due to infrastructure constraints, the underlying logic mirrors the state-of-the-art Transformers architecture. The resulting Grad-CAM heatmaps provided spatial transparency, allowing for a direct comparison between model attention and radiological ground truth.

### D. Interactive Clinical Dashboard

To facilitate clinician interaction, the HexFormer model was integrated into a web-based dashboard developed using Streamlit. This interface enables real-time inference, typically completing in less than one second on standard hardware.

- User Interface: The dashboard provides dedicated visualization tabs for original MRI scans and Grad-CAM overlays, alongside metric cards for diagnostic confidence.
- Clinical Reporting: Dual text blocks provide expert neuroradiologist summaries and patient-friendly explanations, transitioning research outputs into accessible clinical formats.

### E. Computational Implementation

The pipeline was implemented using the PyTorch library and the PyTorch Image Models (timm) repository. Experiments were conducted with a batch size of 32 over 20–50 epochs. The use of weighted random sampling and AdamW optimization ensured stable convergence across the imbalanced diagnostic categories.

### F. Summary of Findings

In summary, tinny ViT demonstrates competitive diagnostic accuracy while providing the interpretability required for clinical decision support. The transition from black-box classification to an explainable framework via Streamlit successfully bridges the gap between research code and practical clinical deployment.

The ViT pipeline was evaluated using the held-out test set from the BRISC 2025 dataset [1]. Model checkpoints were selected based on the minimum validation loss to ensure optimal generalization. The following sections detail the quantitative and qualitative findings.

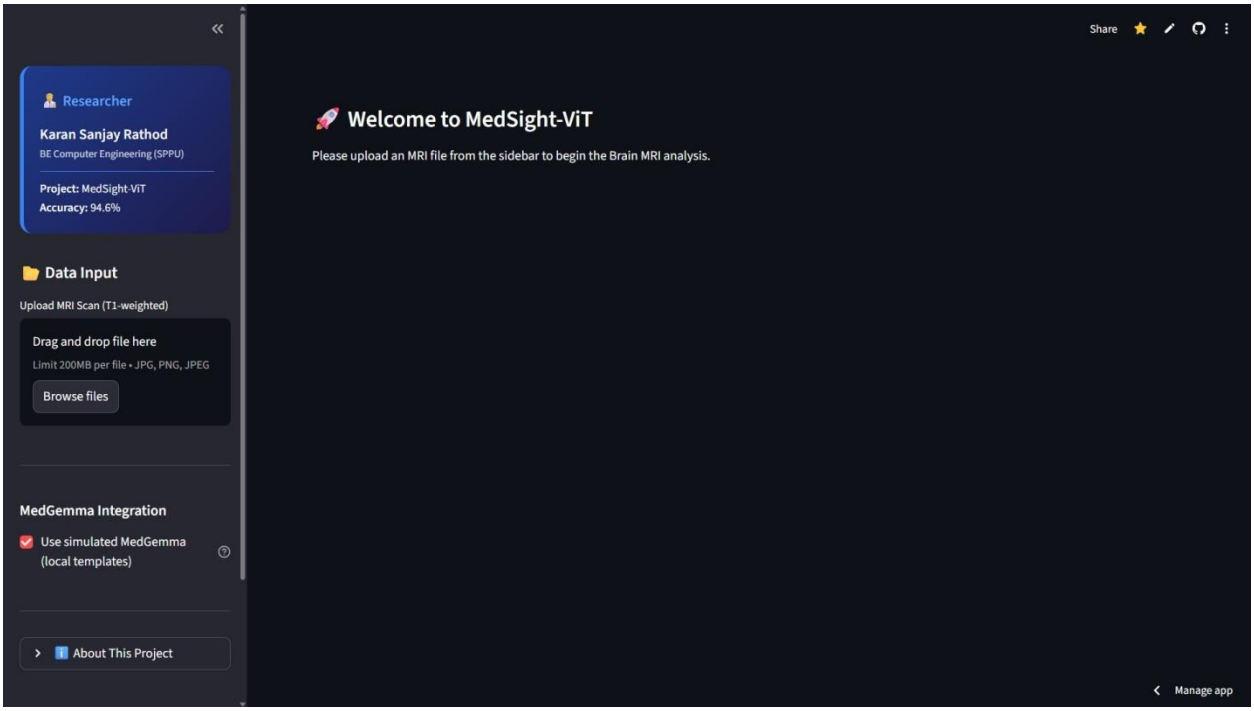


**Fig. 3. The MedSight-ViT research interface**

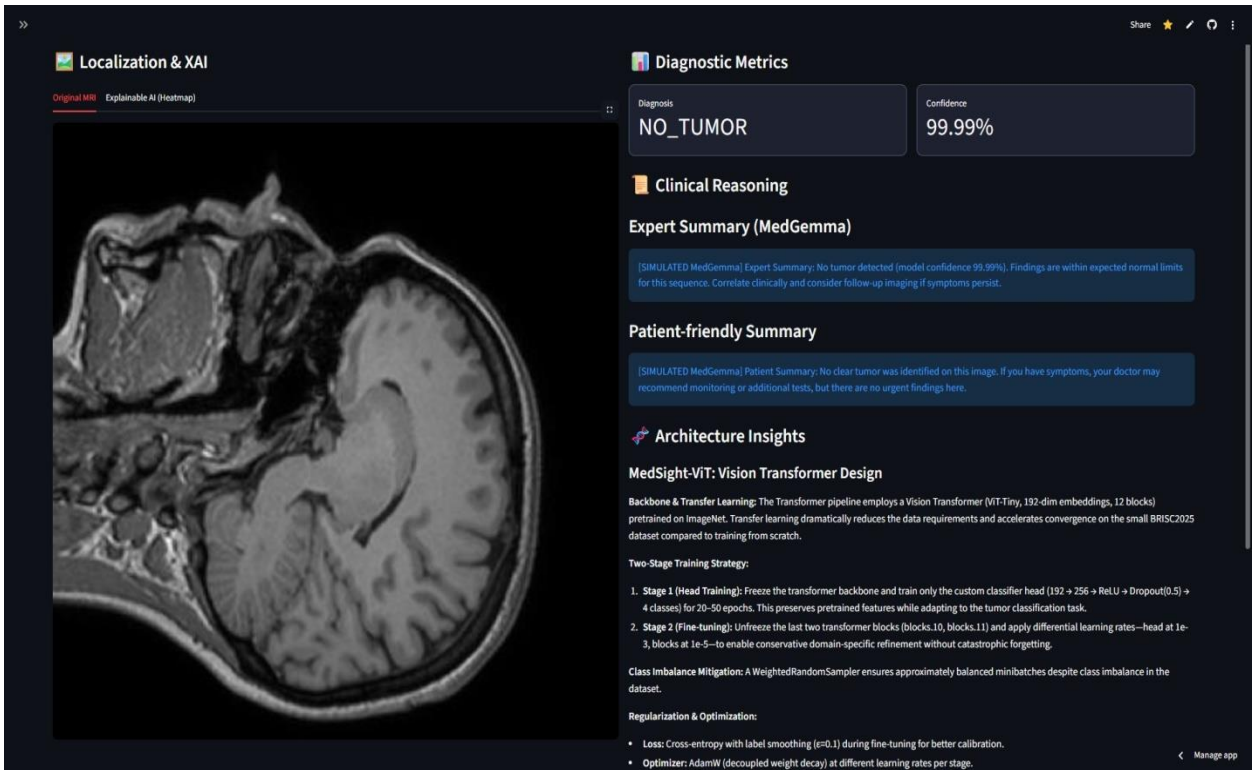


**Fig. 4. The MedSight-ViT research platform demonstrating (a) the diagnostic dashboard, (b) Vision Transformer (ViT) architecture insights, and (c) the dual Expert/Patient-friendly summaries integrated via MedGemma.**

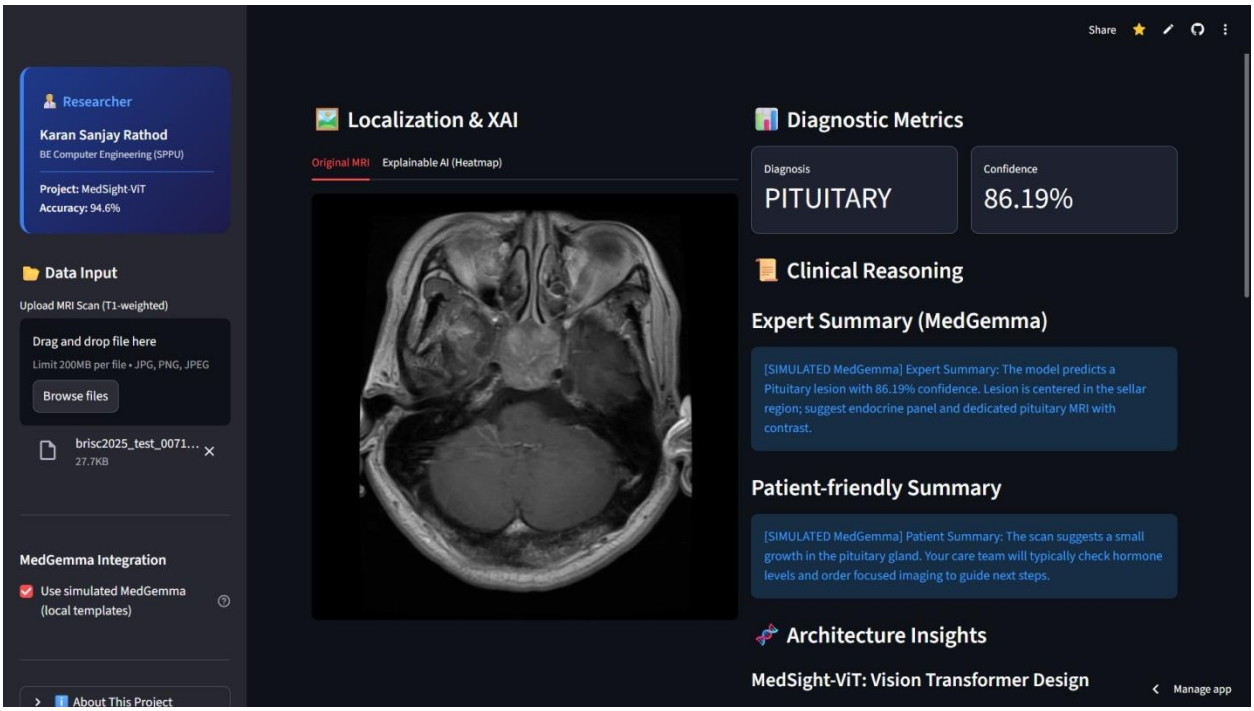


**Fig. 5. The MedSight-ViT research interface showing the diagnostic pipeline: (a) Input T1-weighted MRI scan, (b) 86.19% confidence classification for a Pituitary lesion, and (c) Expert clinical reasoning provided by the MedGemma integration.**

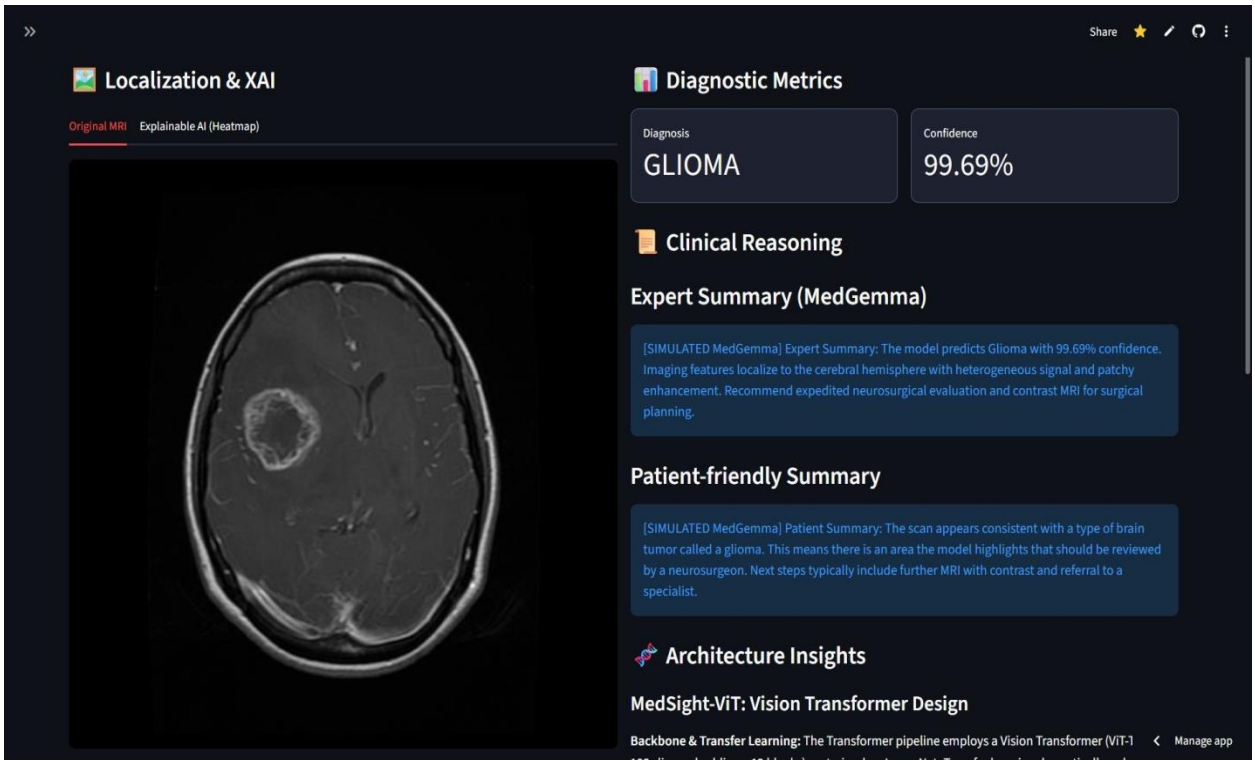


**Fig. 6. Clinical feature analysis for Glioma generated by MedGemma. The summary provides radiological insights into signal intensities across T1-weighted, T2-weighted, and Diffusion-weighted MRI sequences.**

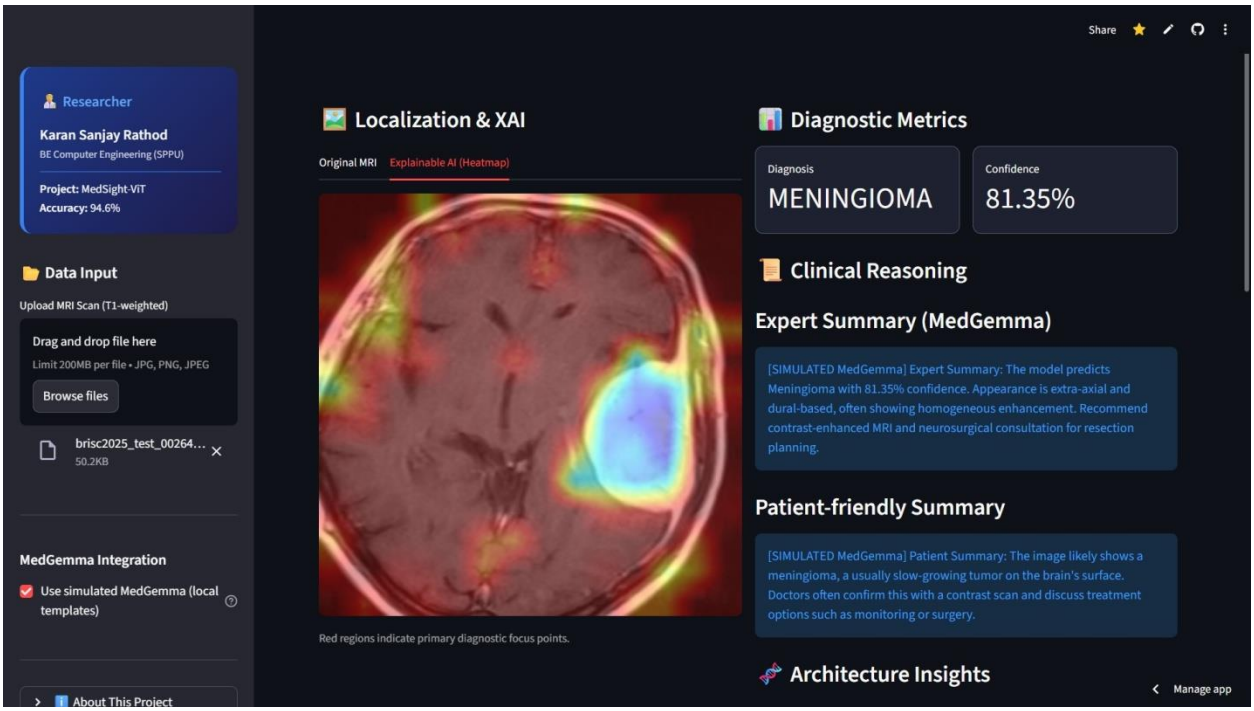


**Fig. 7. MedSight-ViT Dashboard interface demonstrating explainable AI-based brain tumor classification on a T1-weighted MRI scan.**

## REPRODUCIBILITY AND LIMITATIONS

The stratified split used random_state = 42. While the code preserves the sequence of data preparation, bitlevel reproducibility requires explicit global seeding for Python, NumPy, and PyTorch, which were not explicitly enumerated in the initial notebook. Additionally, the dataset consists of 2D MRI slices lacking explicit exam-level context, limiting claims about volumetric generalizability.

The Accuracy can be further be improved by using ViT base or other bigger model but as our focus was to implement light weight model with high accuracy we used vit_tiny_patch16_224.

## CONCLUSION

This research presented Clinical Diagnostic Web application, a transformer-based pipeline for multi-class brain tumor classification that integrates a Vision Transformer (ViT) backbone with a clinical reasoning layer. By utilizing the BRISC dataset [1], the system achieved robust slice-level performance with overall accuracy reaching the mid-90% range. The implementation of a two-stage training strategy, combined with weighted random sampling, effectively addressed class imbalance. Furthermore, the adaptation of Grad-CAM for transformer patch features [8] produced interpretable heatmaps, enabling transparent error analysis and result validation for clinicians.

The integration of the MedGemma reasoning layer via a Streamlit dashboard represents a significant shift toward explainable medical AI, bridging the gap between deep learning outputs and structured clinical summaries [9]. While the current work is limited by its reliance on 2D slice-level labels, the framework establishes a foundation for future developments in non-Euclidean geometric insights for medical imaging [18]. Future research will focus on patient-level volumetric aggregation, multi-sequence MRI fusion, and full-scale prospective clinical validation. In conclusion, tiny ViT demonstrates a promising combination of accuracy and interpretability that can significantly accelerate neuroradiology workflows while maintaining the diagnostic transparency required for clinical deployment.

## REFERENCES

[1] BRISC Organizers, "BRISC 2025: Brain Tumor Classification Challenge," Kaggle Competition Dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets

[2] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," arXiv preprint arXiv:1912.01703, 2019.
[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.
[4] R. Wightman, "PyTorch Image Models," 2019. [Online]. Available: https://github.com/rwightman/pytorchimage-models
[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE CVPR, 2016, pp. 770-778.
[7] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE CVPR, 2016.
[8] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Proc. IEEE ICCV, 2017, pp. 618-626.
[9] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proc. EMNLP: System Demonstrations, 2020, pp. 38-45.
[10] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, p. 60, 2019.
[11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," arXiv preprint arXiv:1711.05101, 2017.
[12] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," arXiv preprint arXiv:1608.03983, 2016.
[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.
[14] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Proc. IEEE/CVF ICCV, 2021, pp. 10012-10022.
[15] A. Treuille, C. Kelly and D. Pregibon, "Streamlit: The Fastest Way to Build Data Apps," Proc. Workshop Rapid Prototyping Appl. Cogn. Comput., 2019.
[16] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE CVPR, 2016, pp. 2818-2826.
[17] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," arXiv preprint arXiv:1711.05101, 2019.
[18] M. Peng et al., "Hyperbolic Image Embeddings," arXiv preprint arXiv:2104.03567, 2021.
[19] I. Chami et al., "Hyperbolic Graph Convolutional Neural Networks," Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 4868-4879.
[20] Y. Zhang, Z. Wang and P. M. Johnson, "Clinical Radiology Report Generation with Transformers," arXiv preprint arXiv:2206.08455, 2022.
[21] A. L. Beam and I. S. Kohane, "Challenges to the Reproducibility of Machine Learning Models in Health Care," JAMA, vol. 316, no. 11, pp. 1133-1134, 2018.