Brain Tumor Detection

Manahil Aamir
24441

Noor us Sabah
25173

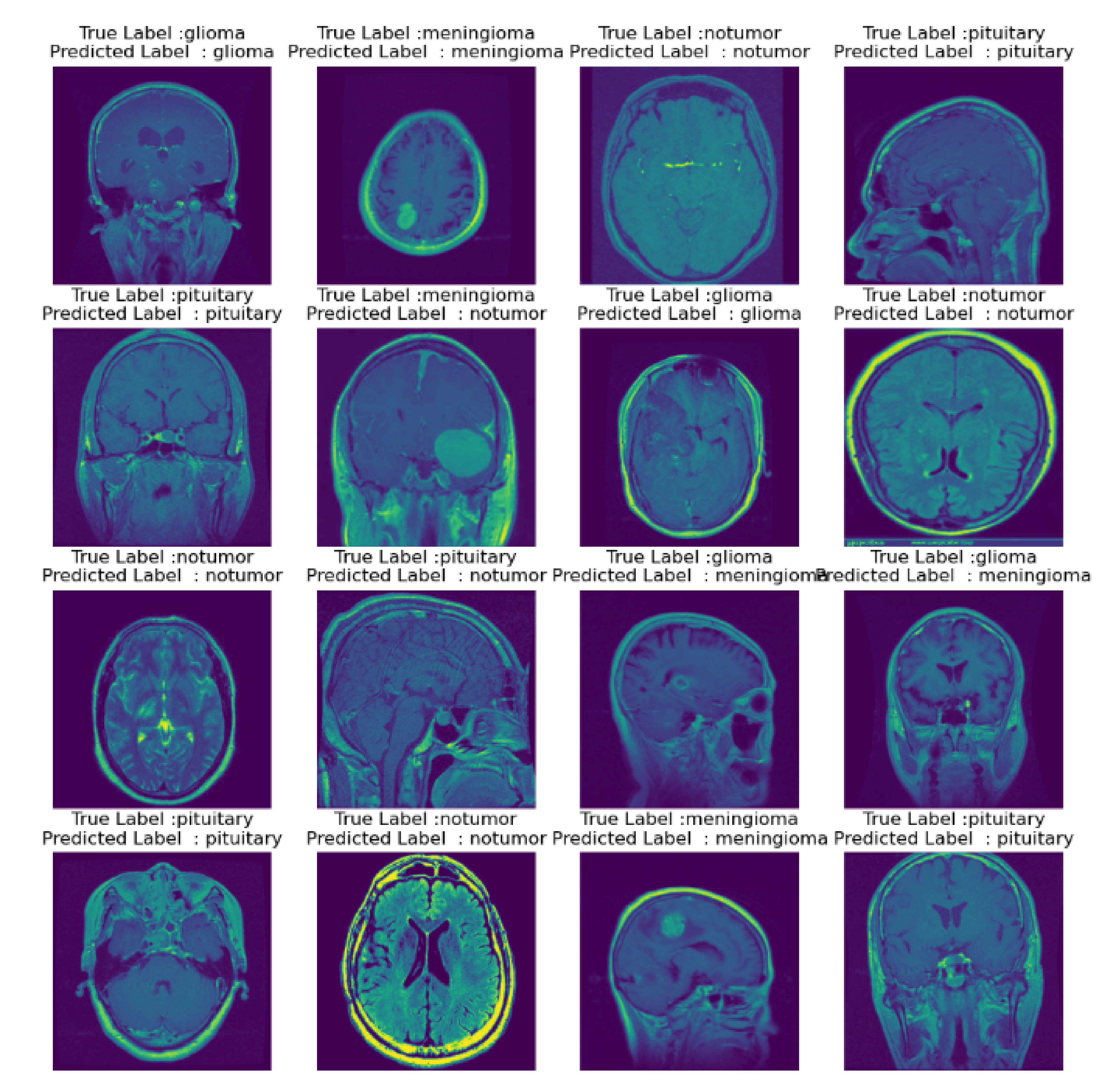
Computer Vision Project

Computer Vision in Diagnosis

Computer vision detects brain tumors by analyzing MRI/CT scans for abnormal patterns using AI. It compares images with past data to spot tumors quickly and accurately. This speeds up diagnosis, reduces errors, and enables early treatment, improving patient outcomes. Automation makes the process efficient and reliable.

CHALLENGE - Each image in the dataset presents a unique challenge due to varying sizes, resolutions, and contrasts.

GOAL - Develop a robust classification model that can accurately identify the presence of brain tumors in MRI scan.



Research | Open access | Published: 23 January 2023

MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques

Soheila Saeedi, Sorayya Rezayi, Hamidreza Keshavarz & Sharareh R. Niakan Kalhori

BMC Medical Informatics and Decision Making **23**, Article number: 16 (2023) | Cite this article

Brain Tumor Detection Using Convolutional Neural Network

Publisher: IEEE Cite This PDF

Tonmoy Hossain ; Fairuz Shadmani Shishir ; Mohsena Ashraf ; MD Abdullah Al Nasim ; Faisal Muhammad Shah All A

Enhanced Brain Tumor Detection Using Integrated CNN-ViT Framework: A Novel Approach for High-Precision Medical Imaging Analysis

Publisher: IEEE Cite This PDF

Safa Jraba ; Mohamed Elleuch ; Hela Ltifi ; Monji Kherallah **All Authors**

Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh

Md. Mahfuz Ahmed, Md. Maruf Hossain, Md. Rakibul Islam, Md. Shahin Ali, Abdullah Al Noman Nafi, Md. Faisal Ahmed, Kazi Mowdud Ahmed, Md. Sipon Miah, Md. Mahbubur Rahman, Mingbo Niu & Md. Khairul Islam

Scientific Reports **14**, Article number: 22797 (2024) | Cite this article

Vision Transformer

- Deep learning model that applies the Transformer architecture to images.
- Splits images into fixed-size patches (like words in NLP)
- Uses self-attention instead of convolutions
- Learns global patterns across the image

Vision Transformer Components

Image Patching and Embedding

- Split - Divide image into fixed-size patches (e.g., 16×16).
- Flatten - Reshape each patch into a 1D vector.
- Embed - Project patches into a higher-dimensional space using a learnable linear layer.

Positional Encoding

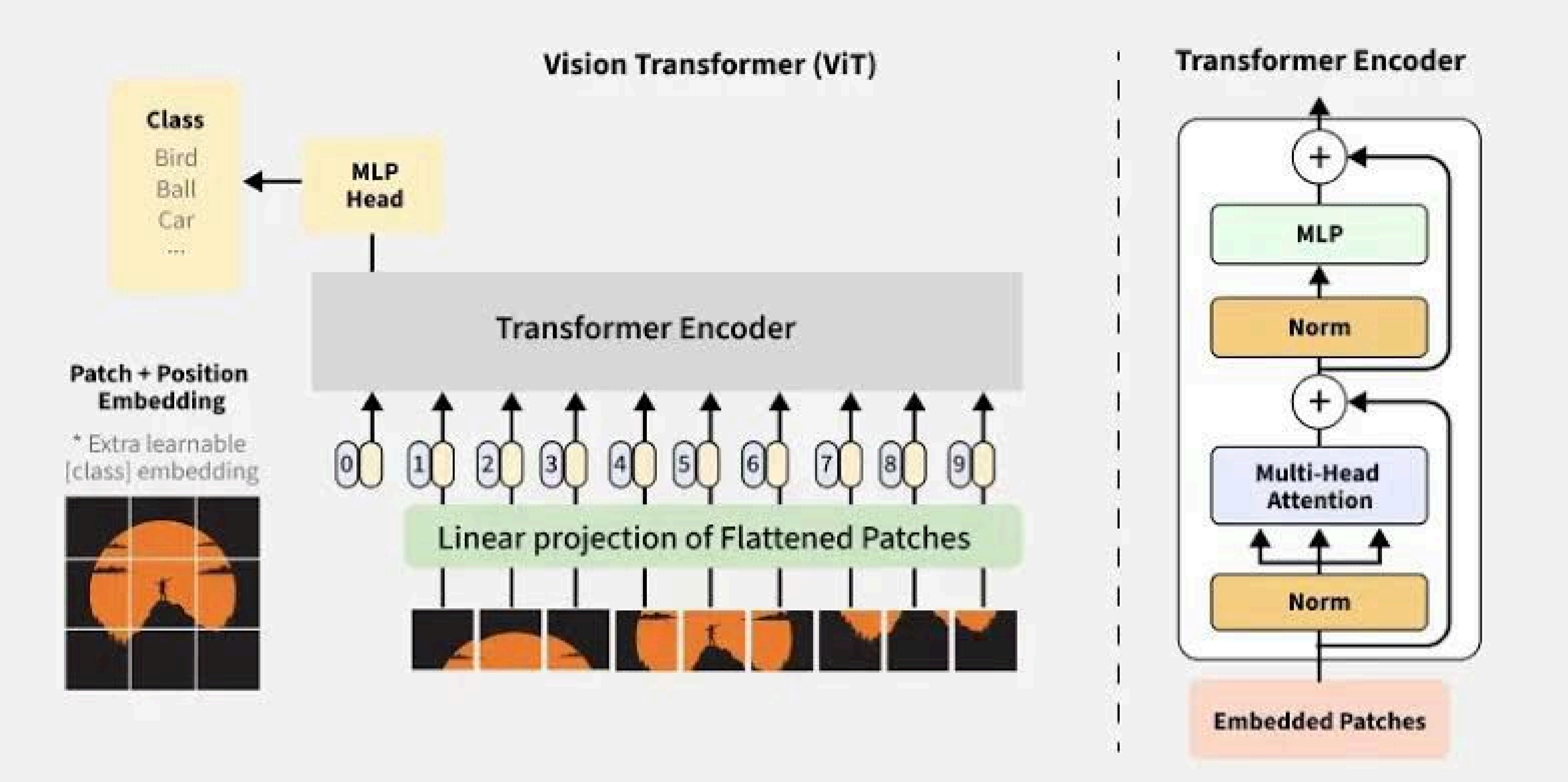
- Positional Embedding - Added to patch embeddings to retain spatial structure

Transformer Encoder

- Multi-Head Self-Attention - Each patch attends to all others to capture global context.
- Feed-Forward Network - Applies a small neural network to each patch for feature refinement.

Classification Head (MLP Head)

- CLS Token - for final prediction after processing by MLP.
- MLP Head - class probabilities via softmax.


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0
1
2
3
4
5
6
7
8
9
Linear projection of Flattened Patches
Transformer Encoder
MLP
Norm
Multi-Head
Attention
Norm
Embedded Patches

Comparative Analysis of ViT Techniques for Tumor Detection

No.	Technique	Accuracy	Training Time
01	ResNet Backbone	0.7788	1 hr 1 mins
02	MobileNet Backbone	0.8459	1 hr 5 mins
03	Patch Encoding	0.8581	1 hr 20 mins
04	Transformer Archiecture Changes	0.9687	1 hr 13 mins
05	Convolutional Stem	0.9725	1 hr 09 mins

Comparison of Input Embedding Techniques

Convolutional Stemming

Downsamples progressively via conv layers (stride-2) for multi-scale features.

Reshapes final feature map into tokens

Preserves local structure better than patch encoding (conv inductive bias).

More efficient & accurate—especially for high-res images.

Patch Encoding

Extracts fixed-size patches (e.g., 16x16 pixels) directly from the image.

Projects patches linearly into a higher-dimensional space (like in ViT).

Adds positional embeddings to retain spatial information.

Simple but rigid—lacks hierarchical feature learning.

Findings of Two Backbone Architectures

ResNet

ResNet needs diverse training samples to generalize well.

If trained on a small dataset, it memorizes noise instead of learning patterns leading to overfitting.

Considerable reduction in accuracy.

Requires RGB pictures and MRI generally grayscale.

MobNet

Uses Depthwise Separable Convolutions

Reduces computation cost significantly (compared to standard CNNs like ResNet).

Slight increase in accuracy

Transformer Encoder Architectural Changes

- Normalizes inputs before attention and FFN (Pre-LN), improving gradient flow.
- Includes dropout layers (attention_dropout, ffn_dropout) for regularization, critical for training stability.
- Uses a 4× expansion in the feed-forward network (default intermediate_size = 4 * hidden_size), following standard Transformer designs (e.g., original "Attention is All You Need" paper).

Thank you!

Happy to Answer Questions!