Brain ViT

Multi-label Image Classification of Brain Pathologies using Vision Transformer (ViT)

- Sainath Vaddi (101179915)
- Neerajdattu Dudam (101179017)
- Sony Reddy Gurram (101179182)
- Drake Michael Farrokhi (A00536349)

Outline

1. Abstract
2. Introduction to Transformers
3. How Vision Transformer (ViT) works?
4. ViT performance in image classification
5. Results and discussion
6. Impacts, and Future work
7. Q&A

Abstract

- Accurate classification of brain pathologies is essential for diagnosing neurological diseases, particularly when co-occurring disorders present subtle distinctions. Multi-label categorization from medical imaging holds promises for facilitating precise diagnoses in such complex scenarios. However, accurately identifying brain abnormalities remains a critical challenge. In this work, we propose BrainViT, a novel solution to address these challenges through simultaneous multi-label classification of brain pathology using Vision Transformers (ViT), which identifies various brain abnormalities accurately from medical images. The proposed model uses a vision transformer that exploits the advantages of the self-attention mechanism, eliminating convolution operations commonly found in traditional deep-learning models for disease detection. Notably, the model's capability to simultaneously identify various brain pathologies in a single pass distinguishes it from conventional methods, providing a more holistic understanding of complex clinical scenarios. The datasets which we want to employ includes tumors like Pituitary, Glioma, Meningioma, and No Tumor. This selection covers a comprehensive range of brain abnormalities. Our proposed BrainViT model represents a significant advancement in the multi-label classification of brain pathologies and contributes to the improvement of accurate and efficient diagnosis in the field of neuroimaging.

Transformers

Why Transformers?

- The Traditional models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) struggled with capturing long-range dependencies in sequential data, due to their sequential processing nature.

- Transformers were introduced in 2017 through the groundbreaking paper titled "Attention is All You Need" by Vaswani et al.

What are Transformers?

- Transformers are a class of deep learning architectures that have gained popular attention, particularly in Natural Language Processing (NLP) tasks. They differ from traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs) in their architecture and mechanisms for processing sequential data.

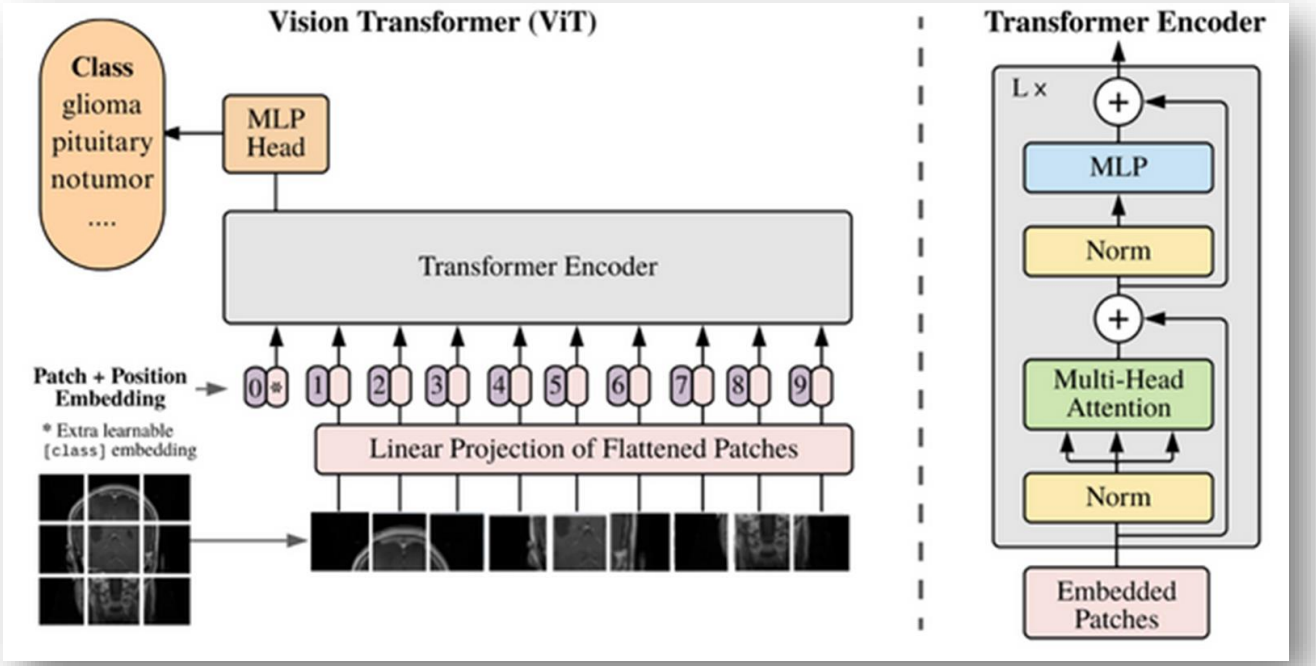
Key Components of Transformers:

- **Self-Attention Mechanism**: Transformers have self-attention mechanism, which enables them to weigh the importance of different elements (called tokens) in a sequence when making predictions. This mechanism allows the model to capture long-range dependencies and relationships between elements, regardless of their positions in the input sequence.

- **Transformer Blocks**: Transformers consist of multiple layers of transformer blocks, each containing self-attention layers and feed-forward neural networks. These transformer blocks enable the model to learn hierarchical representations of the input data, capturing both local and global patterns.

Transition to Vision Transformer

Success of transformers in Natural Language Processing (NLP), led to explore and adapt the transformer architecture to process visual data.



Vision Transformers

- A Vision Transformer is a deep learning architecture that applies the Transformer model, originally developed for NLP, to computer vision tasks.

- **How ViT Works**
- ViT performs on images by breaking them into smaller, fixed-size patches, treating each patch as a token in a sequence.
- This approach allows ViT to apply the same self-attention mechanism used in natural language processing, enabling it to capture spatial relationships between patches.

Components of ViT

Patch Embeddings:

- Images are divided into fixed-size patches.
- These patch embeddings serve as the input tokens for the transformer model.

Positional Encodings:

- Positional encodings help the model understand the relative positions of patches within the image.

Transformer Encoder Layers:

- ViT consists of multiple transformer encoder layers, each containing self-attention mechanisms and feed-forward neural networks.
- These transformer layers enable ViT to learn hierarchical representations of the image, capturing both local and global patterns.

Self-Attention Mechanism:

- Each patch embedding attends to all other patch embeddings, capturing global dependencies within the input image.
- Attention scores are computed between pairs of patch embeddings, determining the importance of each patch with respect to others.
- The attention mechanism enables the model to focus on relevant image regions and learn contextual relationships between patches.

Feedforward Neural Network (FFNN):

- After self-attention, the output representations from each patch are passed through a feedforward neural network (FFNN) independently.
- The FFNN applies non-linear transformations to the features extracted by the self-attention mechanism, enhancing the model's capacity to learn complex patterns and representations.

Layer Normalization:

- Layer normalization is applied after each sub-layer (self-attention and FFNN) within the transformer encoder layers.
- By normalizing input vectors, layer normalization ensures that the model's outputs are centered around zero mean and have unit variance, which can help stabilize the training.

Classification Head:

- At the end of the transformer architecture, a classification head is typically appended to perform specific tasks such as image classification.
- The output representations from the final transformer layer are aggregated and fed into the classification head, which maps them to the output space corresponding to the task.

Training Process

- **Data Preparation:**
- **Dataset Selection:** We started by curating a dataset of brain images annotated with labels corresponding to different brain diseases.
- **Data Preprocessing:** Preprocessing steps included image resizing, augmentation to enhance model generalization and robustness.

- **Model Selection and Architecture Design:**

- **Vision Transformer (ViT):** We chose the Vision Transformer (Pre-trained) architecture for its ability to capture long-range dependencies and learn representations directly from raw image data.
- **Customization:** The ViT architecture was adapted to suit the requirements of our multi-label brain disease classification task. This included adjusting input dimensions, modifying the classification head, and fine-tuning pre-trained model.
- **Loss Function:** For multi-label classification, we utilized binary cross-entropy loss function, which penalizes the model based on the log loss of prediction.

- **Model Training:**

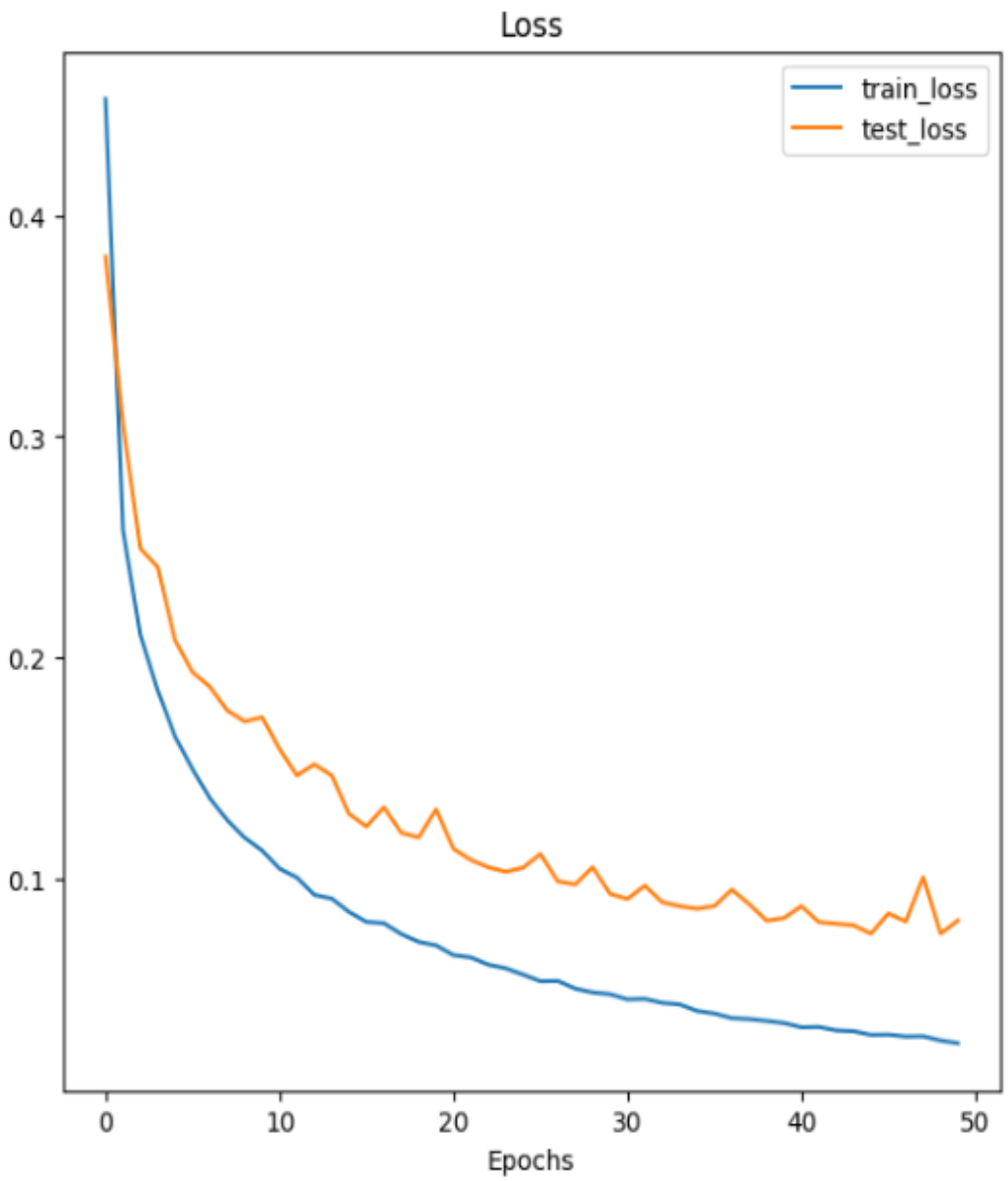
- **Initialization:** We initialized the parameters of the Vision Transformer model using pre-trained weights.

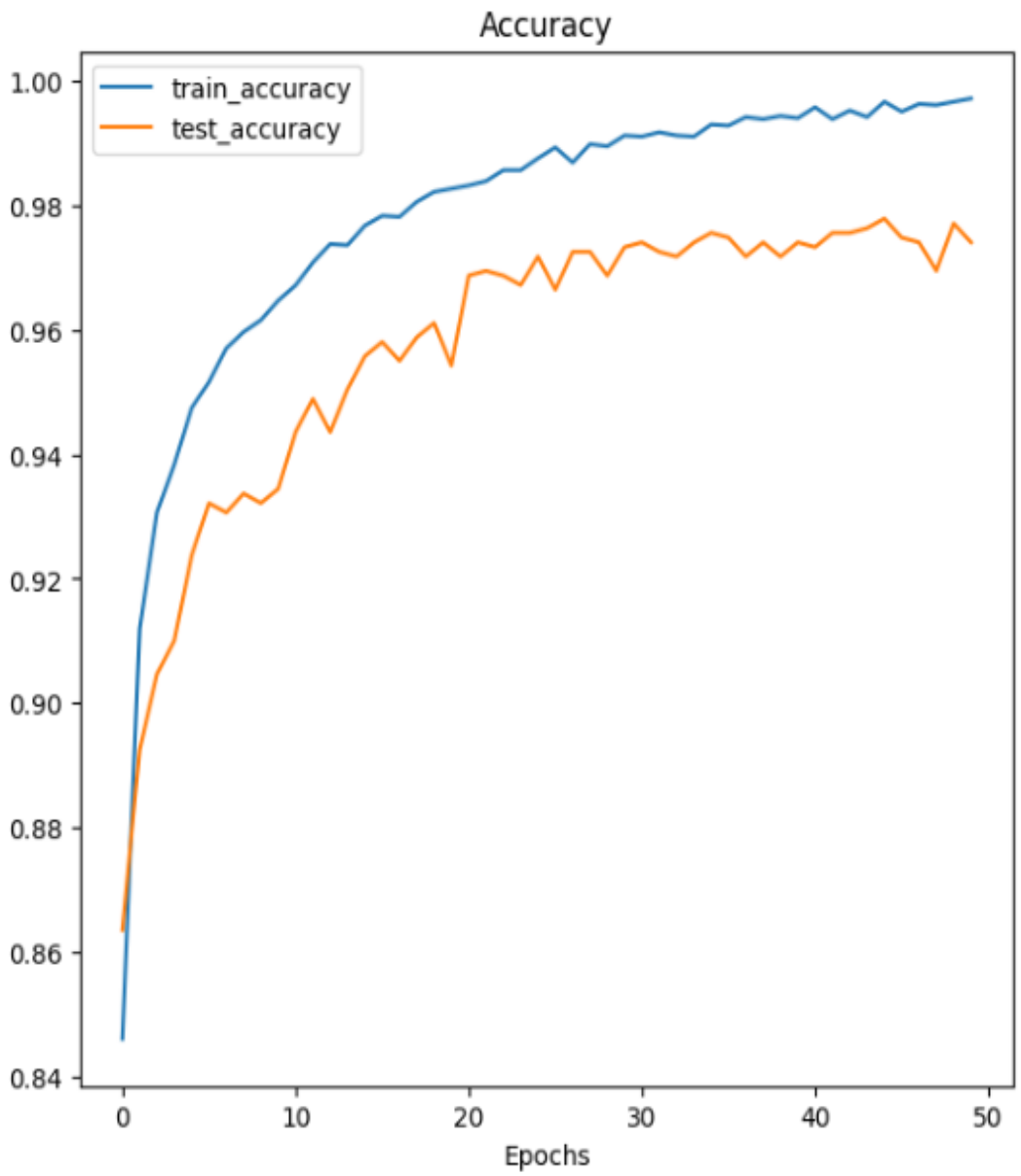
- **Optimization:** Training was performed using optimization algorithms such as Adam, with hyperparameters tuned through experimentation.

- **Training Pipeline:** It involved iteratively feeding batches of brain images and corresponding labels into the model, computing the loss, and updating the model parameters through backpropagation.

- **Monitoring and Evaluation:** We monitored training progress using metrics such as loss and accuracy on validation data.

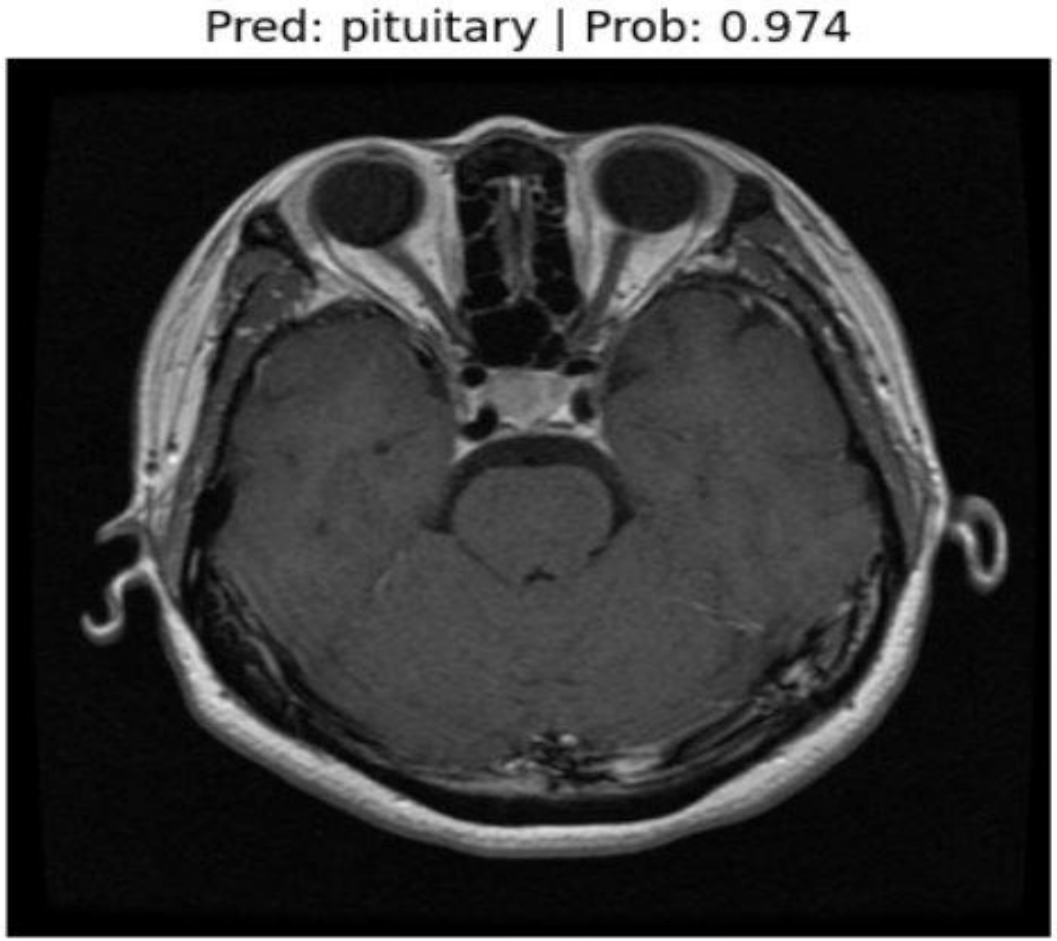
Brain ViT Results



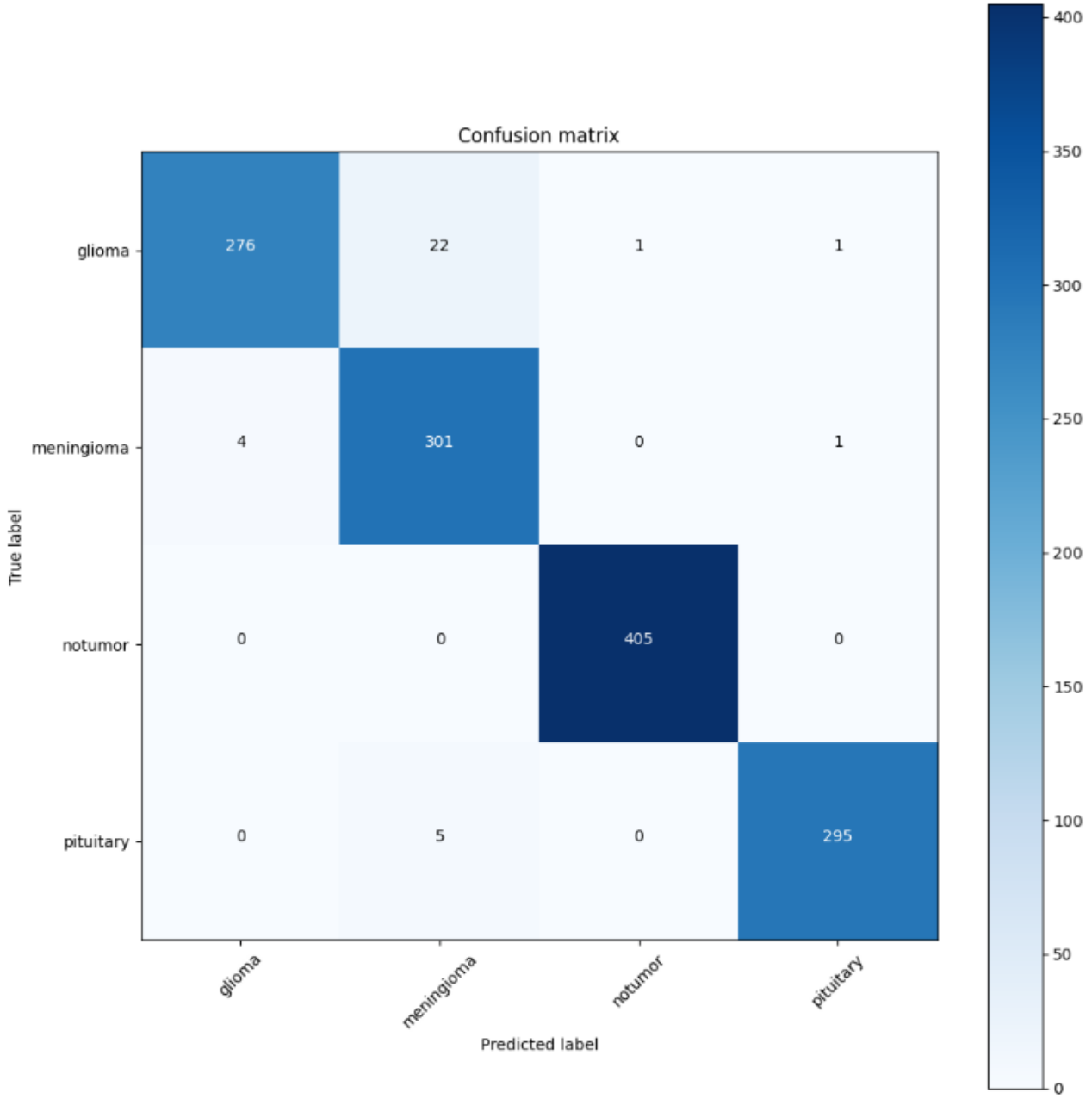


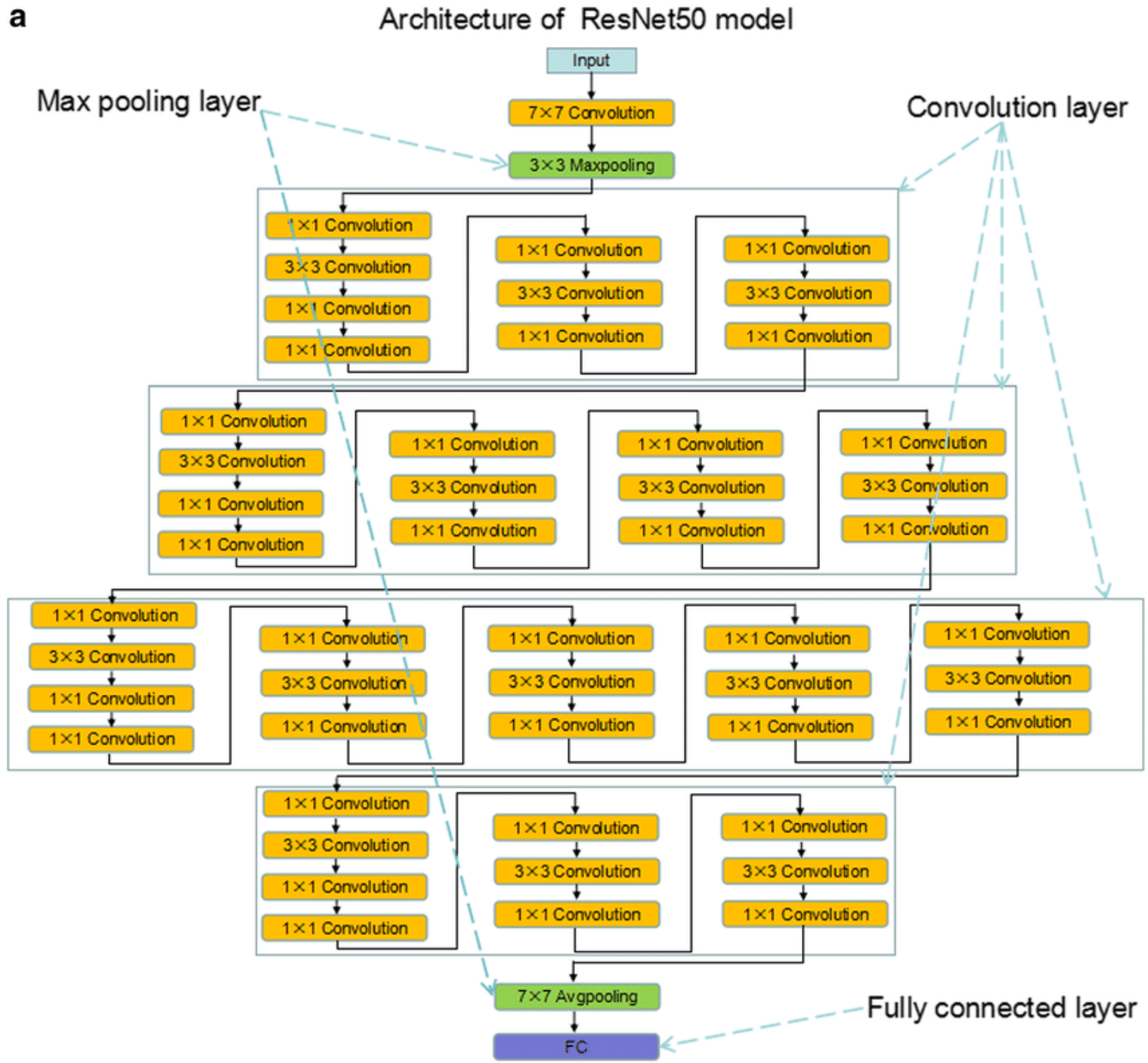
Multi Label Classification Prediction

```
glioma -> 0.007167330477386713
meningioma -> 0.4657747447490692
notumor -> 0.059241138398647 31
pituitary -> 0.9741685390472412
```

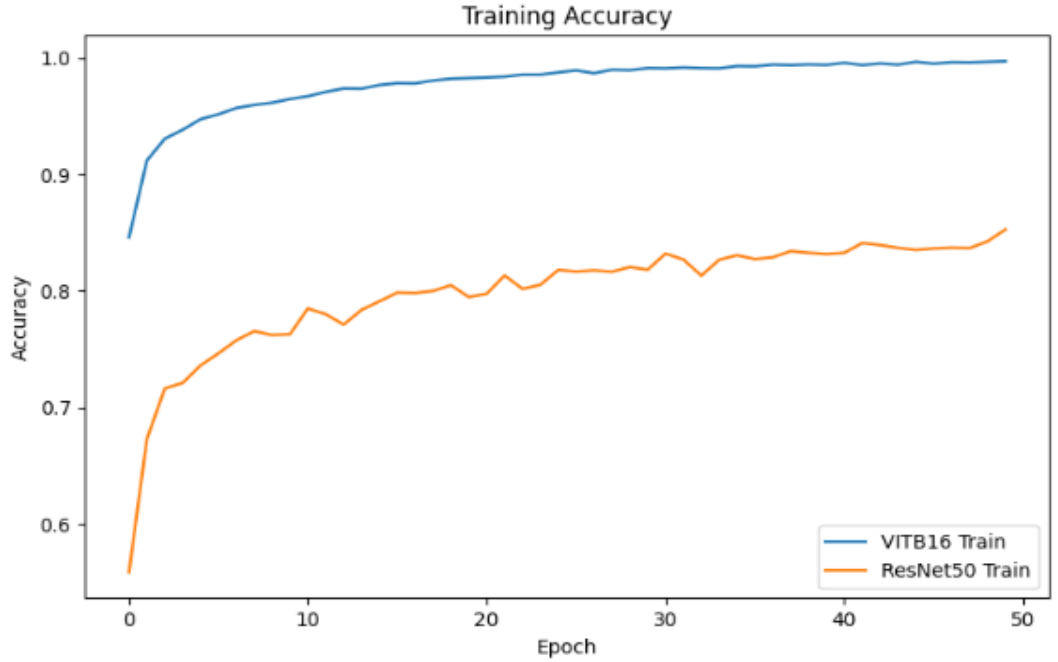


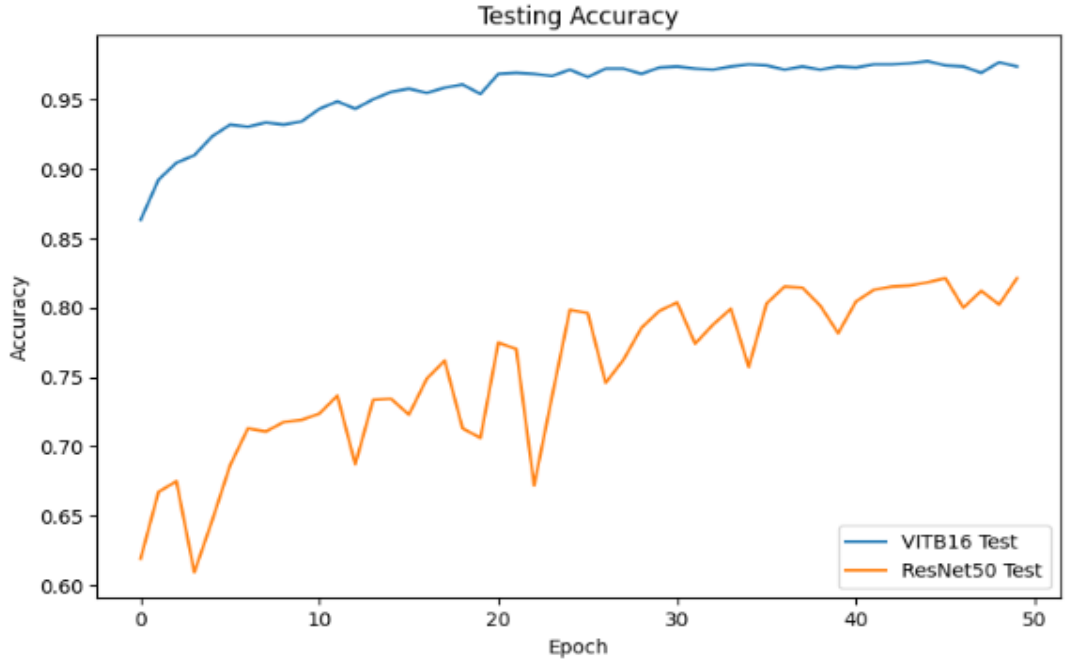
Confusion Matrix

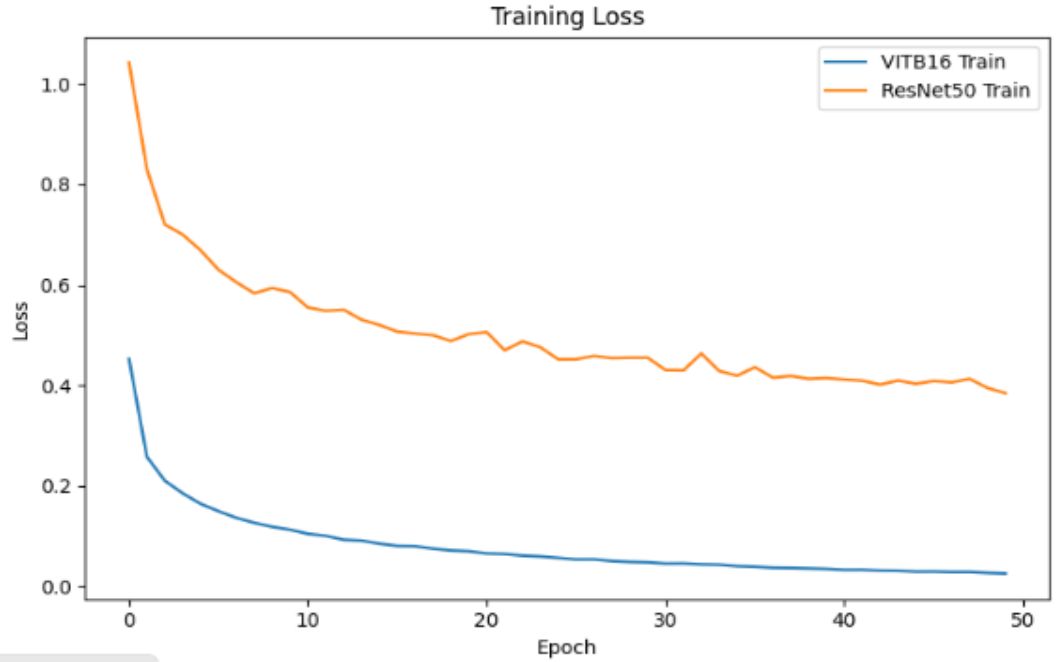


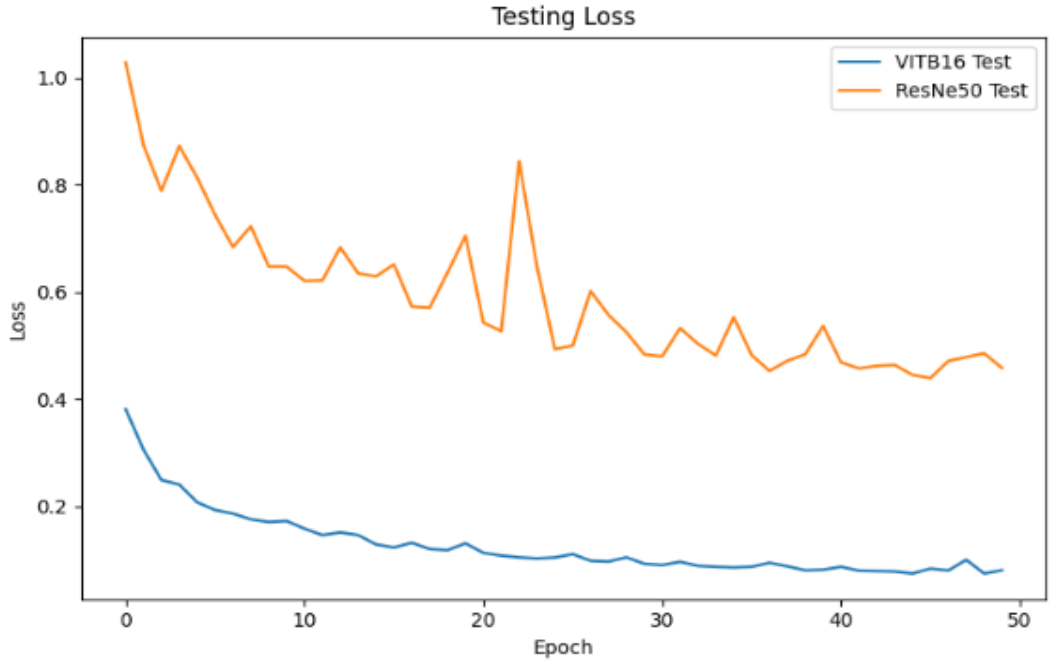

a
Architecture of ResNet50 model
Input
Max pooling layer
7×7 Convolution
Convolution layer
3×3 Maxpooling
1×1 Convolution
3×3 Convolution
1×1 Convolution
7×7 Avgpooling
Fully connected layer
FC

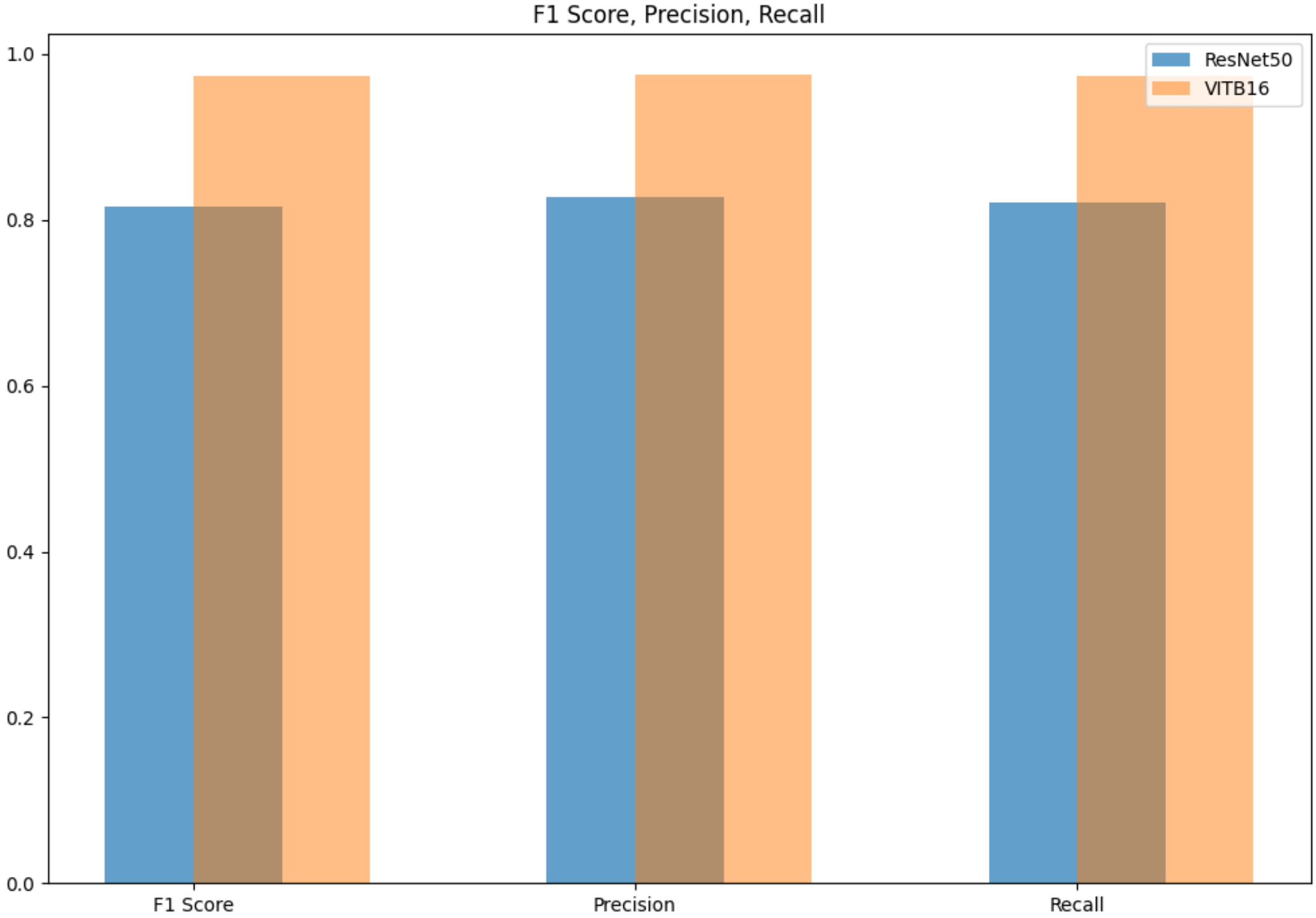
Brain ViT vs ResNet50










F1 Score, Precision, Recall
ResNet50
VITB16
1.0
0.8
0.6
0.4
0.2
0.0
F1 Score
Precision
Recall

ViTB16 outperformed RESNET50

- ViT utilizes self-attention mechanisms that can capture long-range dependencies in the data more effectively than ResNet50 convolutional layers with residual connections, as understanding global context is crucial.
- ViT processes images as sequences of patches and embeds them into tokens for input to the transformer encoder. This tokenization process may have advantages in capturing diverse spatial information across the image compared to ResNet-50,
- Tasks that require understanding relationships between distant parts of an image benefit from models that can capture long-range dependencies well. ViT's self-attention mechanism is designed to handle such scenarios effectively, which might contribute to its superior performance

Advantages over CNN

Convolutional Neural Networks (CNNs)	Vision Transformers (ViTs)
Sequential processing of local receptive fields with shared weights.	Parallel processing of image patches with self-attention mechanism.
Limited ability to capture global context due to local receptive fields.	Capable of capturing long-range dependencies through self-attention mechanism.
Require large number of parameters, especially in deep architectures.	More parameter efficient, achieving similar or better performance with fewer parameters.
Transfer learning in CNNs often requires retraining of many layers.	ViTs facilitate transfer learning with pre-trained models, offering faster adaptation.

Challenges

- Vision transformers often require large amounts of labeled data for training, which may not always be available.

- While vision transformers excel at capturing long-range dependencies, they may struggle with processing very large images due to memory constraints and computational complexity.

- Deployment of vision transformers on resource-constrained mobile platforms or other devices can be challenging due to their computational and memory requirements.

Future Work: Vision Mamba for Brain Disease Classification

- Our future work focuses on vision mamba, an advanced variant of vision transformers, to enhance the accuracy and efficiency of brain disease classification from medical imaging data.

- By leveraging the capabilities of vision mamba, we aim to improve the accuracy and reliability of brain disease classification, enabling more accurate diagnosis and treatment planning.

- We will work towards increasing the diversity and coverage of our brain imaging dataset to encompass a broader range of brain diseases.

Thank You